Measuring the Contributions from a Hydrogen Bond Network to Protein Folding and Coupling in Leucine-Rich Repeat Proteins

by
Sean Anthony Klein

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May, 2019

Abstract

Proteins have evolved to fold cooperatively, though the structural origins of cooperativity are often elusive. Cooperativity is likely to arise from the complex networks of interactions within the long stretches of chemically diverse residues that make up a protein's primary sequence. Unfortunately, the relationship between cooperativity and residue-residue interactions is challenging to study in most systems. Since repeat proteins have a linear structure with a low contact order, they are ideally suited to studies of cooperativity in protein folding. Among repeat proteins, the leucine-rich repeat (LRR) family is of particular interest owing to the presence of a very conserved array of asparagines that may be a source of inter-repeat coupling. This ladder is unique given its high conservation and burial in the hydrophobic core, suggesting it provides enough stabilizing interactions to offset the unfavorable desolvation of its component asparagines. As a linear homogenous network, the asparagine ladder of LRR proteins is a simple system to study the origins of coupling.

This thesis seeks to connect the interactions between asparagine ladder residues to the global properties of stability and cooperativity in LRR proteins. In Chapter 2, a simple two-repeat asparagine ladder in pp32 is studied. The chapter concludes that the asparagine ladder is a stabilizing structural feature that is highly rigid, and reveals interesting difference in hydrogen bonding patterns that make up the ladder. In Chapter 3, a consensus LRR (cLRR) system is evaluated using an extended nearest-neighbor model. The results reveal that, although the A and B repeats have high sequence identity, they have significantly different intrinsic and interfacial terms. In Chapter 4, substitutions

are made to the asparagine ladder in the cLRR series to measure the contributions of ladder asparagines to intrinsic and interfacial stability. Asparagine ladder substitutions are found to affect stability in a directional way, destabilizing N-terminal interfaces more than C-terminal interfaces.

Readers – Prof. Doug Barrick (Advisor) & Prof. Juliette Lecomte

Additional Committee Members – Prof. Bertrand García-Moreno, Prof. Mario Amzel, Prof. Dominique Frueh

Acknowledgements

> We do not deserve our place in the distribution of native endowments, any more than we deserve our initial starting place in society. That we deserve the superior character that enables us to make the effort to cultivate our abilities is also problematic; for such character depends in good part on fortunate family and social circumstances in early life for which we can claim no credit. The notion of desert does not apply here.
>
> – John Rawles, *A Theory of Justice*

Rawles has is it half right but I would extend his assertions. From childhood, through adolescence, and into early adulthood I have been buoyed by the kindness of countless people. At the end of a six year journey, I would be doing them all a disservice to not acknowledge their substantive role in bringing me to this point. Now, with a brief moment to look back I would assert that Rawles should have said our character depends almost wholly on a continuous process of shaping that is wrought by those people chance has thrust upon us and those we've chosen to be a part of our lives.

Chance has been extremely kind to me. If life is like a game then I rolled a natural 20 when I was born to Ron and Michelle Klein (more accurately perhaps, I rolled a natural 4 million). My parents have given me the best possible life I could hope for and supported me in every endeavor with love, wisdom, and plenty of humor. They will forever be at the head of my pantheon, enshrined in their majestic throne of Mortopia. I owe them a debt that can never be repaid, except maybe with grandchildren.

Besides providing me with an endless supply of support, my parents also gave me one of the greatest gifts I will ever receive: my brother Nate. My first and best friend, he will always be a source of inspiration and frustration (and also a kidney, if needed). I would not be where I am or who I am today without him and I am so proud and grateful

to call him my brother. However, scores must be settled and since I am the first brother to publish (at least until his great opus on botany): I did not stab you in the eye with an umbrella, you blocked my umbrella strike into your own eye. Now everyone will know the truth.

Though chance has provided me with wealth that would make Croesus blush, the people that I have met along the road of my life have enriched me beyond my wildest dreams. First among them is Lauren, my wife. I will never know why you settled for me but I will never, ever question it for fear of losing the best thing in my life. Your love got me through the dark times, your joy made the bright times blinding, your piles of clothes will forever decorate our floor. I cannot wait to see where our lives take us and what we will do together. As the twilight princess, you will always get at least 2/3^{rds} of the bed even though you are half the size of me.

While Lauren has been keeping me sane and supporting me outside of lab, I've had an incredible group of people supporting me in lab. First and foremost is my PI, Doug Barrick. He is an incredible teacher and mentor who has given me guidance when I needed it and patience when I needed time and space to grow at my own pace. Of the many, many inspiring things about Doug, the thing that stands out most to me is his passion; if I can find a job that engenders half as much passion in myself as I see daily in him, I will count myself extremely blessed. A testament to Doug's incredible abilities as a PI is the wonderful group he's put together for his lab. Katie Tripp is one of the most fantastic people I've had the pleasure to meet and is a wizard in the lab. Her talents as the head of the Center for Molecular Biophysics are invaluable but, one day, I

hope she finds her way to an idyllic cottage somewhere on the emerald isle. Matt Sternke, Katie's partner in crime, is a steadfast friend and an amazing scientist. A quick wit and a deadly sharp mind will catapult him to the highest echelons of the ivory tower and beyond. Mark Petersen's grit coupled with his intellect and teaching abilities will be an asset to himself and the lab for years to come; I can't wait to see what he accomplishes. Cyril Cook is one of the most patient and persistent people I've ever met (perhaps honed through hours of practice with chess and violin); these traits will pave the way for what I'm sure will be an amazing graduate career. Though I've known Kristen Ramsay for a short time, her drive and smarts are already apparent; I know she will thrive in this lab and use those skills as she moves onward and upward.

Someone else who deserves special mention is Jeliazko Jeliazkov. I must confess, I was minorly worried about moving in with someone who's first and last name are essentially the same. However, that worry morphed into a strong friendship through innumerable games, movies, good jokes (maybe more like a dozen), bad jokes (millions, really), and drinks. From breaking my own phone in a fit of pizza-induced rage to an unforgettable dinner eating General Tso's with copious amounts of beer while watching "Big Trouble in Little China", Jeli has more than earned his place as one of my best friends. I can only imagine what heights he will soar to, buoyed by his limitless energy and impressive mind. I hope that when he gets there, he'll still have time to chat over a beer (or three).

Of course, I would never have had the opportunity to meet most of these people without the mentorship and guidance of Prof. Madeline Shea and her amazing lab

members. She took me in towards the end of my undergraduate career and somehow managed to get me to do more math and analysis at a time when I was trying hard to use my brain for law or surfing my parents couch. Madeline is an incredible mentor who has been a constant source of advice and support, even after I left her lab. I would never have become a biophysicist without her and if she convinced me to do a PhD loaded with scary things (math, physics, computers), I'm pretty sure she can convince anyone of anything (an enviable superpower). Of course, I would never have met Madeline were it not for my steadfast friend, Sterling (sometimes Blake) Martin. His exuberance and passion for science made me question whether I should leave my scientific research behind and certainly played a big role in persuading me to tackle graduate school. Though I am leaving the hallowed halls of university (for now), I know Sterling has many more years ahead of exciting research and I can't wait to see what he can accomplish with his incredible intellect and energy. Another Shea lab alum who has been a big part of my life is Dagan Marx. Though I know I graduated before him, he has always seemed to have the upper hand when it comes to wisdom and knowledge. He has been a sounding board for my ideas, an example to emulate, and a true friend.

So many others deserve thanks: the office staff who always made sure I was taken care of; my friends who've been an equal source of inspiration and fun; our program coordinators who guided me all the way to this very moment; my committee members who gave me guidance and advice, even in the toughest moments; my classmates who were my first friends and kept me sane during the early, hectic years; and so many more people. They deserve more than a line in this acknowledgement and

I hope they can forgive me for my reticence. My time as a graduate student has been truly amazing: new ideas, new skills, friends, mentors, supporters, and even a wife! I can honestly say that this has been one of the best times of my life and I will always look back fondly at what I did an the people who made it possible. As Rawles said, there is no desert here, just a great debt to many, many people. Now, at the very end of a very long journey, my feelings can be summed up by the great Kurt Russel playing R.J. MacReady in John Carpenter's "The Thing":

"I just wanna get up to my shack and get drunk."

Lists of Tables

Lists of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

CHAPTER 1 – Measuring cooperativity in LRR proteins

1.1 Interaction networks in proteins

Cooperativity in any system relies on interactions between residues. The network of interactions that make up cooperativity can be abstracted using graph theory to nodes and edges. For proteins, nodes can represent amino acids with the edges identifying interactions between amino acids. Interestingly, graph representations of proteins adhere to a “small-world” network model, meaning that protein graphs tend to be highly clustered with short path lengths connecting any two residues [1]. As a result, proteins are particularly well suited for allostery [2].

Allostery in its most basic form refers to “action at a distance” [3] arising from exogenous stimulation of a macromolecule. The “small-world” nature of proteins suggests that changes propagate over the short edge lengths in the protein graph [2], which can translate to long distances in real space. Studies of allosteric networks commonly monitor global properties such as stability or binding [2], [4]–[6] with a few studies being able to describe the individual interactions that give rise to the allosteric behavior [7], [8]. However, dissecting perturbations to networks into individual interactions remains exceptionally difficult as these interactions can occur over vastly different time scales and involve large numbers of residues [4]. Detailed descriptions of how interactions propagate in a network are required to better understand natural phenomena like the molecular basis for neurodegenerative disease and improve *de novo* protein design.

A study seeking to provide detailed measurements of the interactions within a network of connected residues would benefit from a system composed of local

interactions between similar types of residues that is amenable to rigorous thermodynamic characterization. Repeat proteins represent a highly simplified architecture ideal for studying interactions between networks of residues. Repeat proteins are composed of repetitions of a single domain or motif to form long arrays with a highly simplified contact order [9]. The linear nature of these proteins dictates that any network of residues that occurs across repeats will itself be linear and will be simplified by the symmetry of the repeat array, compared to the heterogeneous networks present in globular proteins (Figure 1.1). By linearizing the network graph, measuring the effects of substitutions becomes much more tractable and can be accomplished using traditional double-mutant cycles [10]. Another benefit of using repeat proteins to study interaction networks is that nearest-neighbor modeling can be used to quantify the energetics of coupling between neighboring repeats, providing a thermodynamic parameter describing the edges between nodes that connect adjacent repeats [11]. This provides an even more specific measurement of interactions within the network than could be provided by traditional double-mutant cycles.

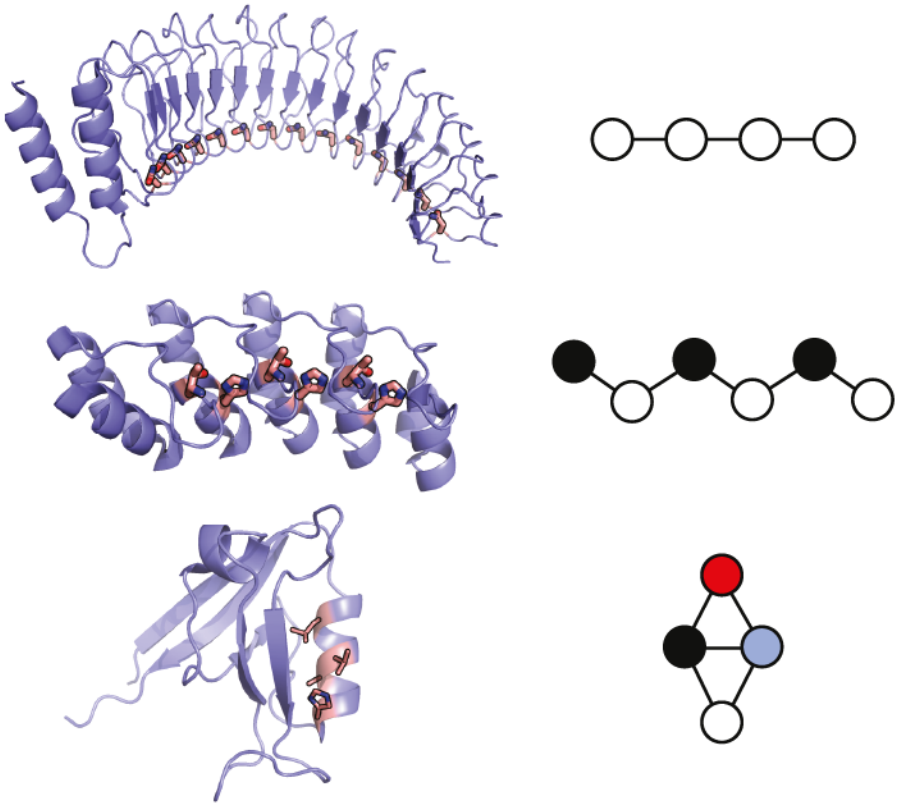


Figure 1.1. Comparison of complexity of residue interactions in repeat and globular proteins. Comparison of residue networks in repeat (top, middle) and globular (bottom) proteins. Structures are shown as ribbon diagrams on the left, with networks of polar residues shown as red sticks. Graphical abstractions of these networks are shown on the right. Nodes represent residues, and different colors indicating different amino acids at each position. The leucine-rich repeat protein YopM (top, PDB ID: 1JL5) has the simplest network architecture, followed by ankyrin (middle, PDB ID: 2BKG), and then PDZ domain (bottom, PDB ID: 1QLC).

1.2 Consensus design and nearest-neighbor modeling

As the nearest-neighbor model has been reviewed extensively elsewhere [12], only a brief description will be provided here. Rather, I will provide a detailed description of how this model is modified to include the effects of substitutions. The power of the nearest-neighbor model is its use of the partition function to fit data rather than an assumption of a two-state model. The partition function for globular proteins is difficult to

construct, since structures are heterogeneous, and the substructures that should be used to represent individual elements are not clearly defined. By comparison, repeat proteins have individual structural units (repeats) clearly defined by symmetry. An even greater simplification can be achieved if all sequence differences between repeats are removed, making the array homogenous and decreasing the number of terms that make up the partition function.

One successful approach to creating homogenous arrays of repeat proteins is to use a consensus sequence for each repeat [13]. In a consensus sequence, the amino acid at each position is determined by the most frequently observed residue at that position in a multiple sequence alignment (MSA). This straight-forward approach can be modified to include charge alternation between repeats; in this approach, the repetitive unit is a pair of repeats that are nearly identical in sequence except for charge alternation [14], [15]. In these cases, poorly conserved positions may be alternated between the top two residues found at the position [16]. Charge alternation is particularly important for β-sheet proteins like those of the leucine-rich repeat family given the close proximity of repeats.

For a homogenous array of identical consensus repeats, only three parameters are needed to construct a 1D nearest-neighbor model: a term to represent folding of an individual repeat (intrinsic free energy, ΔG_R), a term to represent the favorable interaction between repeats (interfacial free energy, $\Delta G_{R\text{-}1,R}$, following the nomenclature of [12]), and the denaturant dependence of the intrinsic term (m_i, Figure 1.2A). This reduction in the

parameters leads to an equally simplified expression for the partition function (with all repeats unfolded as the reference state):

$$q = e^{-2(\Delta G_R - m_i * x)\beta} e^{-\Delta G_{R-1,R}\beta} + 2e^{-(\Delta G_R - m_i * x)\beta} + 1 \qquad (1.1)$$

$$q = \kappa_R^2 \tau_{R-1,R} + 2\kappa_R + 1 \qquad (1.2)$$

$$q = [0 \quad 1] \begin{bmatrix} \kappa_R \tau_{R-1,R} & 1 \\ \kappa_R & 1 \end{bmatrix}^2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad (1.3)$$

where q is the partition function for all states, x is concentration of denaturant (M), β is $(kT)^{-1}$, κ is the equilibrium constant associated with the intrinsic term, and τ is the equilibrium constant associated with the interfacial term. It should be noted that interfaces are treated as denaturant independent (eq 1.1). Interfacial m_i-values have the opposite sign of the intrinsic m_i-values, implying that interfaces become stronger at higher denaturant concentrations. Given the difficulty in explaining this physical phenomenon it has been common practice in many analyses to treat interfaces as being denaturant independent [17]–[19]. Though an interesting (and active) area of research, this thesis will treat interfaces as independent of denaturant as has been done previously.

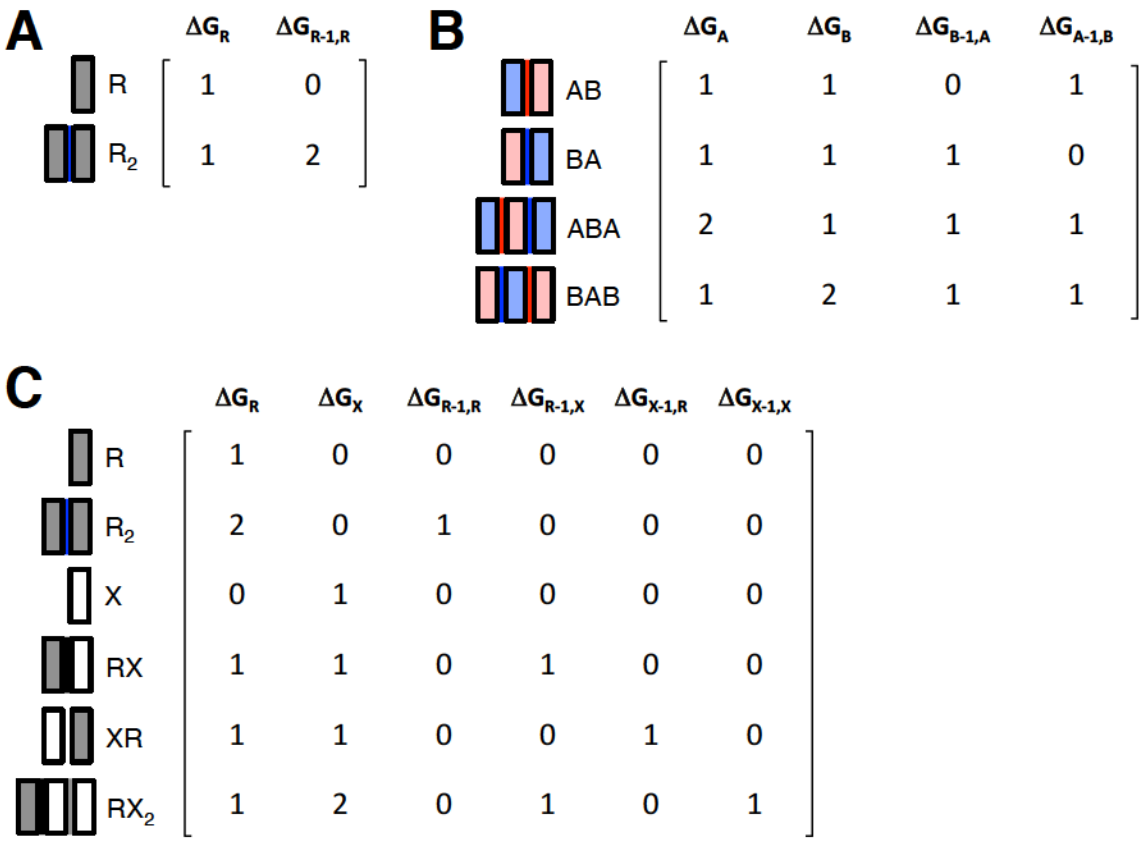


Figure 1.2. Sample matrices for homogenous, alternating, and substituted nearest-neighbor models. Matrices for resolving intrinsic and interfacial parameters for repeat proteins composed of (A) homogeneous repeats, (B) two alternating repeats, and (C) unsubstituted and substituted repeats. Repeats are represented as rectangles with interfaces colored between them and names to the right of the cartoons. Columns are headed by the parameter they represent, and all matrices are full rank. In the homogeneous array, repeats (R) are shown as grey rectangles, and interfaces (R-1, R) are shown with blue shading; only two constructs are needed to resolve the two parameters (ΔG_R and $\Delta G_{R\text{-}1,R}$). In the two-repeat alternating array, the two repeats (A and B) are shown as pink and light blue rectangles, and the two interfaces (A:B and B:A) are shown with red and blue shading; four constructs are needed to resolve the four parameters (ΔG_A, ΔG_B, $\Delta G_{B\text{-}1,A}$, and $\Delta G_{A\text{-}1,B}$). In the substituted repeat array, usubstituted and substituted repeats (R and X) are shown as grey and white rectangles, and interfaces are shown as black, white, and grey shading. Six constructs are needed to resolve the intrinsic (ΔG_X) and interfacial ($\Delta G_{R\text{-}1,X}$, $\Delta G_{X\text{-}1,R}$, $\Delta G_{X\text{-}1,X}$) effects of substitution.

To ensure that parameters are well-determined, different combinations of repeats and interfaces are needed. For the example of a homogenous sequence, each construct has a stability defined as

$$\Delta G^o = \sum_k n_k \Delta G_k + \sum_j n_j \Delta G_{j;i-1,i} \quad (1.4)$$

where k is the different types of repeats in the construct (R in the example), n_k is the number of k repeats in the construct, and j is the types of interfaces in the construct (R-1,R in the example). For any construct, we can represent this linear equation in matrix form:

$$\Delta G^o_{R_2} = [2 \quad 1] \quad (1.5)$$

where ΔG^o_{R2} is the global stability of construct composed of two repeats (R_2), the first column of the matrix represents the number of repeats (2 in the example), and the second column of the matrix represents the number of interfaces. The minimal set of constructs required to resolve a particular set of parameters can be determined from the rank of the coefficient matrix defined by the constructs (Figure 1.2A). A matrix with full rank ensures that all the parameters in the model are independently defined. For the homogenous sequence example two constructs, R and R_2, would be needed to construct the full rank coefficient matrix,

$$\begin{bmatrix} \Delta G^o_R \\ \Delta G^o_{R_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \quad (1.6)$$

using the same format from equation 1.5. With chemical denaturation data from these two constructs and the partition function defined in equation 1.3, we can resolve ΔG_R and

$\Delta G_{R\text{-}1,R}$ by fitting the fraction of each construct that remains folded as a function of urea concentration,

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\frac{q_i}{q} \tag{1.7}$$

where θ is fraction folded, n is the number of repeats, q_i is a sub-partition function that only includes states with the i^{th} repeat folded.

For many consensus systems, homogenous arrays are unstable without solvating caps [18]–[20]. For these systems, interfaces between repeats and caps ($\tau_{N\text{-}1,R}$ or $\tau_{R\text{-}1,C}$) can be treated the same as interfaces between repeats [17]. Under this assumption, the full-rank coefficient matrix for a homogeneous repeat with caps is

$$\begin{bmatrix} \Delta G^o_{NR} \\ \Delta G^o_{NRC} \\ \Delta G^o_{NR_2C} \\ \Delta G^o_{RC} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 3 \\ 0 & 1 & 1 & 1 \end{bmatrix} \tag{1.8}$$

where the columns of the coefficient matrix represent the intrinsic (ΔG_N ,1st; ΔG_R 2nd; ΔG_C 3rd) and interfacial ($\Delta G_{R\text{-}1,R}$, 4th) terms from equation 1.4. If $\Delta G_{N\text{-}1,R}$ and $\Delta G_{R\text{-}1,C}$ are not equal to $\Delta G_{R\text{-}1,R}$, the difference between the actual and assumed interfaces will accrue to the intrinsic terms for the caps (ΔG_N and ΔG_C). Fortunately, ΔG_N and ΔG_C are not of interest in most studies as the caps are present to prevent aggregation, not provide accurate measurements of thermodynamic parameters.

The partition function of capped or uncapped homogeneous arrays can easily be adapted to model heterogeneous arrays, for example a consensus repeat protein with two different repeat sequences (A and B) arranged sequentially (Figure 1.2A) [18]. For

such constructs the intrinsic (and interfacial) terms are expected to differ from one repeat (and interface) to the next. Therefore, two intrinsic and two interfacial terms are required, doubling the total number of constructs needed to resolve the two-repeat system. The partition function for a construct composed of AB repeat pairs is

$$q = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \kappa_A \tau_{B-1,A} & 1 \\ \kappa_A & 1 \end{bmatrix} \begin{bmatrix} \kappa_B \tau_{A-1,B} & 1 \\ \kappa_B & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (1.9)$$

This two-repeat partition function is used to describe the stability of a consensus LRR protein in Chapter 3.

Another useful application of the nearest neighbor approach is in analysis of the effects of one or more substitutions in an otherwise homogenous array of repeats (Figure 1.2C) [20], [21]. This creates a new repeat (X) with a new intrinsic parameter (ΔG_X). Because the substitution may modify either of the adjacent interfaces, two new interfacial parameters are needed, one for the N-terminal interface ($\Delta G_{R\text{-}1,X}$), and one for the C-terminal interface ($\Delta G_{X\text{-}1,R}$). If multiple substitutions are made at adjacent sites, a fourth new parameter ($\Delta G_{X\text{-}1,X}$) is needed to capture nonadditivity between substitutions [10]. The partition function of NRX_2 reflects this additional complexity:

$$q = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \kappa_N \tau_{R-1,R} & 1 \\ \kappa_N & 1 \end{bmatrix} \begin{bmatrix} \kappa_R \tau_{R-1,R} & 1 \\ \kappa_R & 1 \end{bmatrix} \begin{bmatrix} \kappa_X \tau_{R-1,X} & 1 \\ \kappa_X & 1 \end{bmatrix} \begin{bmatrix} \kappa_X \tau_{X-1,X} & 1 \\ \kappa_X & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (1.10)$$

Note that a new matrix is required for each repeat as a result of the different τ term associated with each interface. Fully resolving a substitution using a nearest-neighbor model requires more constructs than using a two-state model, but provides much greater

resolution of the effects of substitutions. This substitution system is used to evaluate the effects of substitutions to a consensus LRR protein in Chapter 4.

What remains to be described is how these systems can be used to resolve the effects of networks of interacting residues. As mentioned previously, the repeat protein architecture greatly simplifies a potential interaction network. If the interaction network is conserved, it will persist in the consensus repeat sequence and can then be studied using mutagenesis coupled with the nearest-neighbor modeling approach described above that includes the effects of substitutions. Analyzing the unfolding transitions of an appropriate set of constructs with a substitution nearest-neighbor model provides detailed information on the network: if the network is within repeats, the intrinsic terms are expected to be affected, whereas if the network spans repeats, the interfacial terms should reflect any perturbations to the network resulting from substitution. As substitution interfaces are described by twice the number of parameters that are associated with intrinsic terms, more information might be available for networks that span repeats than those within repeats. The requirements of the nearest-neighbor analysis have therefore provided a framework for selecting possible systems for study: the system must be a repeat protein, it must have a network of interacting residues, that network must be retained in a consensus sequence repeat array, and the network must span repeats (be interfacial in nature) to maximize the amount of information provided per construct evaluated. One group of proteins that satisfies all these requirements is the leucine-rich repeat (LRR) family, which are the subject of this thesis.

1.3 LRR protein structure

LRR proteins are a diverse family of repeat proteins found in all domains of life [22]. They form curved solenoids with concave faces comprised of parallel β-sheets linked via a turn to a helical or coiled convex face (Figure 1.3) [23]. LRRs participate in a wide variety of binding interactions with diverse biological functions including immunity (TLRs [24], agnathan antibodies [25]), and kinase signaling [26]. Binding interactions generally take place on the β-sheet or coiled region between the concave and convex faces of the LRR solenoid [27].

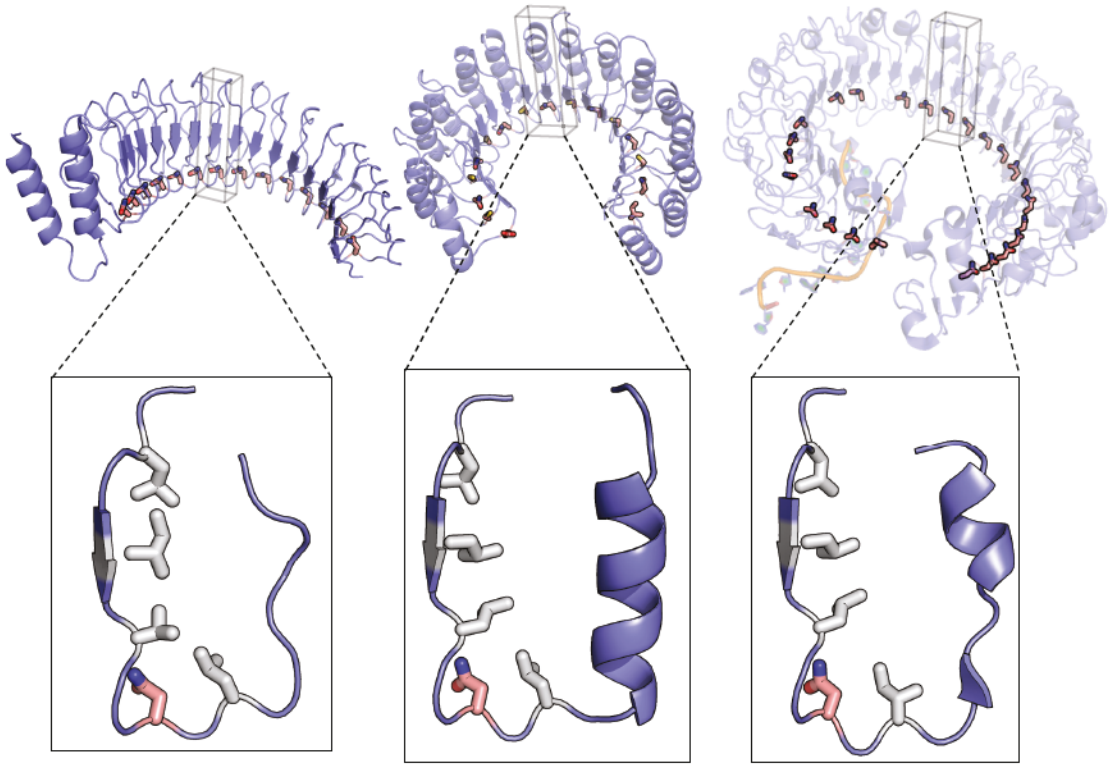


Figure 1.3. Structural diversity in LRR subfamilies. Examples of LRR protein structures (top) and an individual repeat from each structure (bottom). From left to right: YopM (PDB ID: 1JL5) from the bacterial LRR subfamily, ribonuclease inhibitor (PDB ID: 1A4Y) from the ribonuclease inhibitor subfamily, and TLR9 (PDB ID: 3WPC) from the typical subfamily. Asparagine ladder residues are shown as red sticks. In the individual repeats, conserved leucines in the invariant LRR region are shown in white sticks.

The biological diversity of LRRs has produced a significant amount of sequence diversity in the LRR family. LRRs can be categorized into a number of subfamilies that exhibit different lengths and conservation patterns (Figure 1.4) [27]. However, a shared feature of all LRR sequences is the 11-residue "invariant region" (LxxLxL/VxxN/CxL) that forms the β-sheet and part of the succeeding coiled region (Figure 1.3) [28]. As can be seen in Figure 1.4, most of the sequence and length variation arises from the positions C-terminal to the invariant region [27].

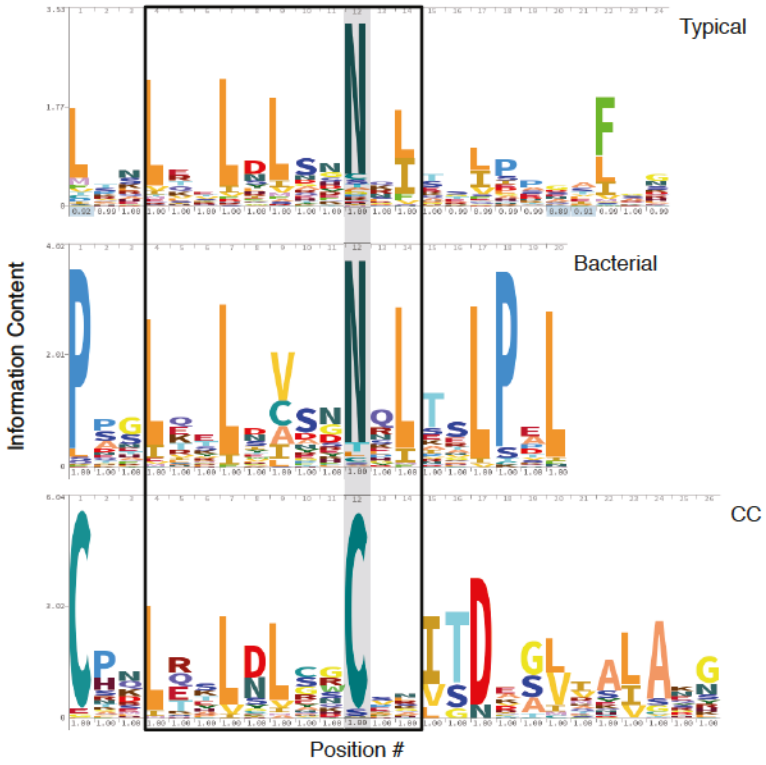


Figure 1.4. Conservation patterns in LRR subfamilies. Comparison of Hidden Markov Model (HMM) logos [29] constructed from typical (top), bacterial (middle), and cysteine-containing (CC, bottom) LRR subfamily sequences. The invariant region is within the black rectangle with the conserved asparagine ladder highlighted with in gray. Numbers below each column indicate the percentage of sequences in the MSA with a residue (as opposed to a gap) at that position. Numbers above columns representing the position number.

For this thesis, the most important structural feature of LRR proteins is the highly conserved asparagine ladder (Figure 1.3 and 1.4) [30]. The asparagine ladder is formed through stacking of multiple LRRs into regular and repeating array, apparently sequestering the asparagine side chains from solvent (Figure 1.3). Similar structural motifs are also found in β-helical proteins and amyloids [31]. The asparagine ladder is particularly well suited for studying networked interactions as it is one-dimensional (due to the repeat protein architecture), homogenous, and may play role in repeat-to-repeat coupling through hydrogen bonding interactions between repeats (Figure 1.5). In short, the asparagine ladder of LRR proteins is an ideal structural motif to study short- and long-range interactions within a network.

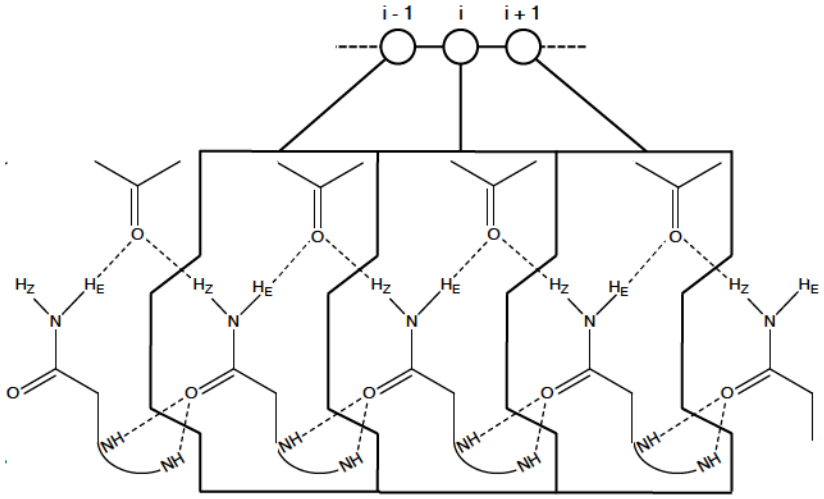


Figure 1.5. Graphic and schematic representation of the asparagine ladder. A graphical abstraction of the asparagine ladder from Figure 1.1 with nodes representing the ladder asparagine in repeats *i - 1*, *i*, and *i + 1* connected to a schematic showing the side chain interactions observed in crystal structures. Dashed lines from the nodes indicate that the pattern can extend to include an arbitrary number of nodes. Dashed lines in the bonding scheme (bottom) represent hydrogen bonds. Labels show asparagine side chain protons (H_Z/H_E) and backbone amides (NH).

In chapter 2, the asparagine ladder in the LRR protein pp32 is studied to determine the global and molecular properties of a simple two-asparagine ladder that is amenable to NMR spectroscopy and thermodynamic analysis. These studies reveal that the asparagine ladder is an important component of global LRR stability, and that it is highly rigid. Chapter 3 lays the foundation for studying the asparagine ladder in a consensus LRR designed by Dr. Thuy Dao [16]. Parameters for a single-repeat system are resolved to provide a baseline from which to measure changes in stability resulting from asparagine ladder substitutions in chapter 4. This final chapter describes substitutions to the cLRR asparagine ladder, using the nearest-neighbor substitution model to analyze unfolding transitions. Both cysteine and leucine substitutions are described and the ladder is found to have polarized stability with substitutions causing effects at non-adjacent interfaces.

1.4 References

[1] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, “Small-world view of the amino acids that play a key role in protein folding,” *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, 2002.

[2] N. V. Dokholyan, “Controlling Allosteric Networks in Proteins,” *Chemical Reviews*. 2016.

[3] J. Monod, J. P. Changeux, and F. Jacob, “Allosteric proteins and cellular control systems,” *J. Mol. Biol.*, 1963.

[4] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, “The ensemble nature of allostery,” *Nature*, vol. 508, no. 7496, pp. 331–339, 2014.

[5] A. S. Raman, K. I. White, and R. Ranganathan, “Origins of Allostery and Evolvability in Proteins: A Case Study,” *Cell*, 2016.

[6] V. H. Salinas and R. Ranganathan, “Coevolution-based inference of amino acid interactions underlying protein function,” *Elife*, 2018.

[7] M. W. Clarkson, S. A. Gilmore, M. H. Edgell, and A. L. Lee, “Dynamic coupling and allosteric behavior in a nonallosteric protein,” *Biochemistry*, vol. 45, no. 25, pp. 7693–7699, 2006.

[8] E. J. Fuentes, S. A. Gilmore, R. V. Mauldin, and A. L. Lee, “Evaluation of Energetic and Dynamic Coupling Networks in a PDZ Domain Protein,” *J. Mol. Biol.*, 2006.

[9] E. Kloss, N. Courtemanche, and D. Barrick, “Repeat-protein folding: New insights into origins of cooperativity, stability, and topology,” *Arch. Biochem. Biophys.*, vol. 469, no. 1, pp. 83–99, 2008.

[10] A. Horovitz, “Double-mutant cycles: A powerful tool for analyzing protein structure and function,” *Fold. Des.*, vol. 1, no. 6, pp. 121–126, 1996.

[11] T. Kajander, A. L. Cortajarena, E. R. G. Main, S. G. J. Mochrie, and L. Regan, “A new folding paradigm for repeat proteins,” *J. Am. Chem. Soc.*, 2005.

[12] T. Aksel and D. Barrick, *Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models*, 1st ed., vol. 455, no. A. Elsevier Inc., 2009.

[13] L. K. Mosavi, D. L. Minor, and Z. -y. Peng, “Consensus-derived structural determinants of the ankyrin repeat motif,” *Proc. Natl. Acad. Sci.*, 2002.

[14] M. T. Stumpp, P. Forrer, H. K. Binz, and A. Plückthun, “Designing repeat proteins: Modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family,” *J. Mol. Biol.*, vol. 332, no. 2, pp. 471–487, 2003.

[15] R. Parker, A. Mercedes-Camacho, and T. Z. Grove, “Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment,” *Protein Sci.*, 2014.

[16] T. P. Dao, A. Majumdar, and D. Barrick, “Capping motifs stabilize the leucine-Rich repeat protein PP32 and rigidify adjacent repeats,” *Protein Sci.*, 2014.

[17] J. D. Marold, J. M. Kavran, G. D. Bowman, and D. Barrick, “A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins,” *Structure*, 2015.

[18] K. Geiger-Schuller and D. Barrick, “Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States,” *Biophys. J.*, vol. 111, no. 11, pp. 2395–2403, 2016.
[19] K. Geiger-Schuller, K. Sforza, M. Yuhas, F. Parmeggiani, D. Baker, and D. Barrick, “Extreme stability in de novo-designed repeat arrays is determined by unusually stable short-range interactions,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 29, p. 201800283, 2018.
[20] T. Aksel and D. Barrick, “Direct observation of parallel folding pathways revealed using a symmetric repeat protein system,” *Biophys. J.*, vol. 107, no. 1, pp. 220–232, 2014.
[21] K. A. Sforza, *Alpha-helical repeat protein folding and turnover: A thermodynamic analysis of natural and unnatural repeat architectures*. Johns Hopkins University, 2017.
[22] J. Bella, K. L. Hindle, P. A. McEwan, and S. C. Lovell, “The leucine-rich repeat structure,” *Cell. Mol. Life Sci.*, vol. 65, no. 15, pp. 2307–2333, 2008.
[23] B. Kobe and J. Deisenhofer, “The leucine-rich repeat: a versatile binding motif,” *Trends in Biochemical Sciences*. 1994.
[24] S. G. S. T. C. Buchanan and N. J. Gay, “Structural and functional diversity in the leucine rich repeat family of proteins.,” *Prog. Biophys. Molec. Biol.*, vol. 65, no. I, pp. 1–44, 1996.
[25] C. A. Velikovsky *et al.*, “Structure of a lamprey variable lymphocyte receptor in complex with a protein antigen,” *Nat. Struct. Mol. Biol.*, 2009.
[26] E. Smakowska-Luzan *et al.*, “An extracellular network of Arabidopsis leucine-rich repeat receptor kinases,” *Nature*, vol. 553, no. 7688, pp. 342–346, 2018.
[27] B. Kobe and A. V. Kajava, “The leucine-rich repeat as a protein recognition motif,” *Current Opinion in Structural Biology*. 2001.
[28] A. V. Kajava, “Structural diversity of leucine-rich repeat proteins,” *J. Mol. Biol.*, vol. 277, no. 3, pp. 519–527, 1998.
[29] T. J. Wheeler, J. Clements, and R. D. Finn, “Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models,” *BMC Bioinformatics*, 2014.
[30] B. Kobe and J. Deisenhofer, “Proteins with leucine-rich repeats,” *Curr. Opin. Struct. Biol.*, 1995.
[31] J. Hennetin, B. Jullian, A. C. Steven, and A. V. Kajava, “Standard Conformations of β-Arches in β-Solenoid Proteins,” *J. Mol. Biol.*, vol. 358, no. 4, pp. 1094–1105, 2006.

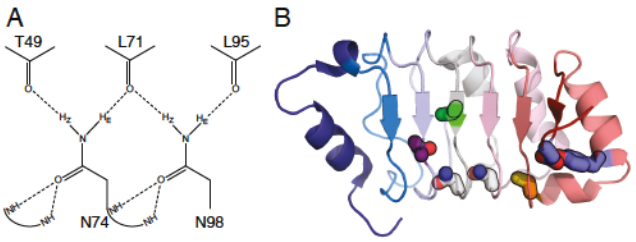
Chapter 2 - A second backbone: the contribution of a buried asparagine ladder to the global and local stability of a leucine-rich repeat protein

2.1 Introduction

Networks of polar interactions in protein interiors are common structural motifs that play important roles in stabilizing proteins [1]–[3]. The unusual environments of these interior polar networks are likely to impart unusual energetic features on the networks, and have long been recognized as contributing to communication processes such as allostery [4]. Determining the contributions of individual residues to folding and/or stabilizing the native states of proteins is not straightforward, as many studies can attest [5]–[7]. An even greater challenge lies in identifying collective networks of residues whose interactions are key to protein cooperativity, allostery, and function. Because these networks are defined by sequence information at multiple positions, understanding interaction networks in proteins often requires a large amount of sequence information [8], [9], accurate sequence alignments [10], [11], sophisticated analyses [12]–[14], and creation and experimental characterization of large numbers of variants [15]–[17] compared to studies focused on single-site contributions.

Although it can be difficult to identify functionally relevant polar residues (specifically, those enhancing stability) and their interaction networks, there are some cases where key residues and interaction networks can be inferred using structural data and sequence conservation alone. One such example is the linear array of asparagine residues, often called an asparagine ladder, found in leucine-rich repeat (LRR) proteins.

Asparagine ladders are highly conserved both in sequence and in structure, and are entirely buried in the hydrophobic interior of LRR proteins despite the polar character of their eponymous asparagine residue (Figure 1 and S1). Compared to other networks of interactions, the LRR asparagine ladder is an attractive target for experimental study since it is composed of the same residue (asparagine) at the same position within adjacent repeats, reflecting a simple symmetry across the network. Furthermore, the asparagine ladder is likely to be tightly coupled via hydrogen bonds (Figure 1A), providing a unique opportunity to study how a highly conserved network contributes to protein stability and cooperativity. A study of the asparagine ladder in LRR proteins may also provide a better understanding of asparagine/glutamine ladder architectures in other contexts, perhaps most notably in the amyloid protein aggregates responsible for neurodegenerative pathologies like Alzheimer's (Aβ), Parkinson's (α-synuclein), and Huntington's (huntingtin) diseases [18].



C

Segment	Sequence	Residues
LRR Motif	LxxLxLxxNxL	
N-Cap	MEMGRRIHLELRNRTPSD	1-18
Repeat 1	VKELVLDNSRSNEGK-LEGLTDE-FEE	19-43
Repeat 2	LEFLSTINVGLTS---IANLPK--LNK	44-65
Repeat 3	LKKLELSDNRVSGG--LEVLAEK-CPN	66-89
Repeat 4	LTHLNLSGNKIKDLSTIEPLKK--LEN	90-114
Repeat 5	LKSLDLFNCEVTNLNDYRENVFKLLPQ	115-141
C-Cap	LTYLDGYDRDDKE	142-154

Figure 2.1. The structure and sequence of pp32 and its asparagine ladder. (A) The hydrogen bond network of the side chain NH_2 groups of asparagines 74 and 98. Potential hydrogen bonds to backbone carbonyls identified in the crystal structure are shown as dashed lines. (B) Structure of pp32 (PDB ID: 4XOS) with substituted residues shown as sticks. Repeats are colored from blue (N-terminus) to red (C-terminus). Key residues in this study are colored as follows: T49, purple; N74 and N98, gray; L69, green; C123, orange; Y131 and D146, blue. (C) The canonical LRR motif and the pp32 sequence separated into caps and repeats. Residues are aligned based on the HMM logo in Figure S1; minor shifts in the alignment were introduced based on the three-dimensional structure of pp32. Positions corresponding to the LRR motif are highlighted gray.

In this study, we examine the role of the asparagine ladder in LRR folding and evaluate the behavior of its hydrogen bond network. Using the LRR protein pp32, we determined the thermodynamic contribution of the asparagine ladder to global stability and cooperativity using urea-induced unfolding monitored by circular dichroism. To explore the structural and dynamic features of the hydrogen bond network in pp32, we used NMR spectroscopy to measure chemical shifts, ^{15}N dynamics of the backbone and asparagine side chains, temperature coefficients, and hydrogen exchange rates of asparagine ladder side chains compared to solvent-exposed asparagine side-chains. Our results show that the asparagine ladder plays an important role in stabilizing the LRR fold, and its disruption causes substantial changes to secondary structure and loss of stability.

Furthermore, thermodynamic parameters indicate the asparagine ladder contributes to more cooperativity than the conserved leucine residues that define the LRR motif [19]–[21]. NMR chemical shifts and temperature coefficients suggest that hydrogen bonds formed by NH_2 groups of ladder asparagine side chains are stronger between repeats than they are within repeats. These data indicate the asparagine ladder acts like a second backbone, maintaining the LRR fold and supporting repeat-to-repeat interactions that promote global cooperativity.

2.2 Results

2.2.1 The structure and sequence features of asparagine ladders in LRR proteins.

In the asparagine ladder motif of LRR proteins, the primary amide side chains of asparagines are completely buried within the nonpolar environment of the protein core. Owing to the high hydrogen bonding capacity of primary amides and their capacity to donate and accept hydrogen bonds equally (donating two hydrogen bonds from the NH_2 group and accepting two hydrogen bonds to the carbonyl oxygen), these buried asparagine side chains are likely to form extended networks of hydrogen bonds. Crystal structures of LRR proteins show that this is accomplished through hydrogen bonds to the amide groups of peptide bonds (Figure 1A), wherein two peptide amide NH groups donate hydrogen bonds to each asparagine $C_\gamma O$ group, and two peptide carbonyl oxygens accept hydrogen bonds from each asparagine NH_2 group. Though other polar and charged protein side-chains (serine, threonine, histidine, arginine, aspartate and glutamate) have high hydrogen bonding capacities, there is a clear preference for asparagine at the ladder

position of typical LRR repeats (84 percent of repeats, Figure S1). Only cysteine, serine, and threonine occur at the ladder position with frequencies greater than 1 percent (9.3, 1.5, and 1.4 percent), implying that the stereochemistry and hydrogen-bond patterning of the asparagine side-chain is well matched for the ladder environment.

Owing to the repetitive architecture of LRR proteins, analogous peptide groups from each repeat participate in hydrogen bonding to each buried asparagine. Three of these hydrogen bonds cross the interfaces between adjacent repeats. Specifically, the backbone NH donor groups to each asparagine $C_\gamma O_{\delta 1}$ and one of the carbonyl acceptor groups to each asparagine NH_Z are from the previous repeat (residue *i-24* and *i-22* for donors, residue *i-27* for acceptor; Figure 1A). The fourth hydrogen bond is between the asparagine ladder NH_E and the backbone CO of residue *i-3* in the same repeat; in addition, this backbone CO also acts as the *i-27* donor to the asparagine in the neighboring C-terminal repeat. Therefore, this network of backbone carbonyl acceptors forms a continuous hydrogen bond network ($O_{i-27}...H_ZN_iH_E...O_{i-3}...H_ZN_{i+24}H_E$) that spans the entire ladder. It seems likely that this network contributes to strong coupling of adjacent repeats.

Because the ladder asparagine is one of the most conserved residues in some LRR families (Figure S1), the length of these ladders can span large distances. For example, the asparagine ladder in YopM spans 14 continuous repeats [22]. Although the length of this extended hydrogen bond network is impressive, a large protein like YopM would be a challenging target to apply high-resolution studies of hydrogen bonding, such as NMR spectroscopy. Thus, for the present study, we sought to identify a simple

asparagine ladder in a relatively small LRR protein. The LRR domain from the human pp32 protein meets this criterion, with two ladder asparagines centered in a short array of LRRs (Figure 1B). Though internal asparagines in longer ladders may have different properties than the two-residue ladder in pp32, the shared structural features of all asparagine ladder residues (most notably, their hydrogen bonding patterns) suggests their salient properties are preserved in the short pp32 ladder [19], [23]. Furthermore, the central ladder position in pp32 also permits us to determine the effects of ladder extending substitutions in adjacent repeats. pp32 is also well-suited for the present studies because high-resolution structures have been determined [24], [25] and its thermodynamic stability and folding kinetics have been well characterized [21], [26], [27].

pp32 has five canonical LRRs flanked by N- and C-terminal caps (Figure 1B). As pp32 and its homologues are found predominantly in the animal taxon and have repeat lengths ranging from 21 to 26 residues, the LRRs of pp32are best represented by the "typical" LRR subfamily [19], [28]. The asparagine ladder in pp32 is composed of the side chains of asparagines 74 (repeat three) and 98 (repeat four). Unlike repeats three and four, residues at the analogous positions in repeats one and two are hydrophobic. The residue at the ladder position in repeat five is a cysteine, the second most common residue at the ladder position (Figure 1C, S1).

2.2.2 The effect of asparagine ladder substitutions on pp32 global stability.

To determine the importance of asparagines 74 and 98 to the stability and structural integrity of pp32, we generated a series of constructs in which asparagines 74 and/or 98 are substituted with alanine and leucine. With the exception of the N98A variant,

far-UV CD spectra of these constructs differ significantly from the wild-type pp32 spectrum (Figure 2A), indicating that the secondary structure is perturbed by the loss of ladder asparagines. The increase in negative ellipticity from ~200 to 210 nm suggests that substitutions of asparagines 74 and 98 increase the amount of random coil present in pp32 for all variants except N98A. Nonetheless, all spectra retain a pronounced shoulder near 218nm, suggesting that some amount of the native β-sheet structure is preserved.

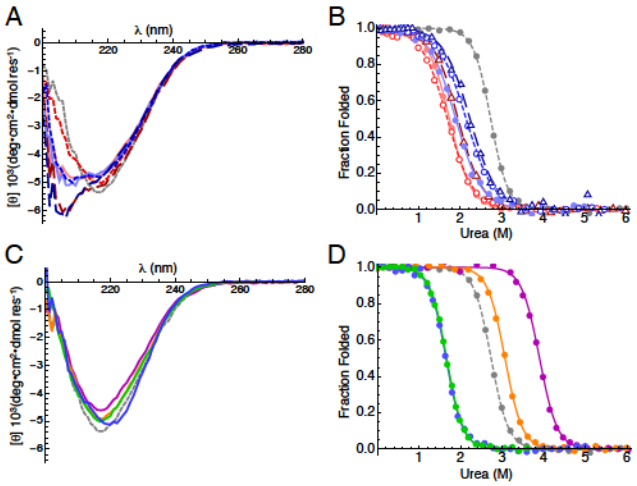


Figure 2.2 CD spectra and urea-induced unfolding of asparagine ladder and peripheral variants of pp32. (A) Far-UV CD spectra of asparagine ladder variants. Alanine variants are colored red (N74A, solid line; N98A, short dashed line; N74A/N98A, long dashed line), leucine variants are colored blue (N74L, solid line; N98L, short dashed line; N74L/N98L, long dashed line), and wild-type pp32 is colored grey (dashed line). (B) Urea melts of asparagine ladder variants. Transitions were monitored by CD at 220 nm and were fitted using a two-state model (curves). Data and curves are transformed to fraction folded. Colors and line styles are as in (A). (C) Far-UV CD spectra of peripheral variants (T49L, purple; L69A, green; C123N, orange; YD, blue). (D) Urea melts of peripheral variants. Data were collected and transformed as in (B). Colors and line styles are as in (C). Raw titration data are shown in Figure S2. Conditions: 20 mM $NaPO_4$, 150 mM NaCl, 0.1 mM TCEP, pH 7.8, 20 °C.

To determine how hydrophobic substitutions to the asparagine ladder impact folding stability and cooperativity, we collected urea-induced unfolding transitions using CD spectroscopy (Figure 2B, S2A). For all variants, substitution of ladder asparagines with hydrophobic residues is significantly destabilizing (Table 1, average $\Delta\Delta G^{\circ}_{H2O}$ = 4.3 kcal mol^{-1}) and reduces the steepness of the unfolding transitions (Table 1, average decrease in m-value = 1 kcal mol^{-1} M^{-1}). This decrease in the m-value suggests a decreased cooperativity in unfolding, though the reduced m-value may also result from partial disruption of structure in the absence of denaturant, consistent with the loss of secondary structure observed by far-UV CD. However, the N98A variant shows a significant decrease in cooperativity (Δm = 1.0 kcal mol^{-1} M^{-1}) but has a far-UV spectrum that is the same as wild-type pp32 (Figure 1A), implying that the m-value is due to a genuine decrease in cooperativity rather than partial unfolding under native conditions.

Table 1. Global stability for pp32 variants.

Variant	ΔG°_{H2O}[a]	m-value[a]
wild-type[b]	-7.93 ± 0.18	2.86 ± 0.02
S27N	-5.45 ± 0.18	2.66 ± 0.07
T49L	-10.49 ± 0.33	2.67 ± 0.06
T49V	-9.76 ± 0.42	2.76 ± 0.13
V52N	-5.62 ± 0.4	2.77 ± 0.20
T49L/V52N[c]	-9.35	2.80
C123N	-8.90 ± 0.23	2.87 ± 0.07
L69A[b]	-4.99 ± 0.13	3.02 ± 0.03
YD[b]	-4.72 ± 0.14	2.69 ± 0.14
N74A	-3.28 ± 0.22	1.98 ± 0.11
N98A	-3.18 ± 0.14	1.85 ± 0.06
N74A/N98A	-3.62 ± 0.25	1.94 ± 0.06
N74L	-3.61 ± 0.03	1.94 ± 0.02
N98L	-3.68 ± 0.2	1.59 ± 0.11
N74L/N98L	-4.18 ± 0.35	1.83 ± 0.12

[a]Global stabilities were determined from urea-induced unfolding transitions at 20 °C. Units for ΔG°_{H2O} and m-values are kcal mol^{-1} and kcal mol^{-1} M_{urea}^{-1}. Uncertainties are standard deviations on the mean from at least three independent unfolding transitions. [b]Equilibrium unfolding data are from [21]. [c]Only a single measurement was made so no error is reported.

This observation demonstrates that the asparagine ladder plays a role in promoting cooperative folding, in addition to stabilizing the LRR fold. It is noteworthy that substitution of both ladder residues with hydrophobic residues (for example, the doubly-substituted N74L/N98L variant) appears to be no more destabilizing than single substitution (the singly substituted N74L and N98L variants; Figure 2B, Table 1). Rather, comparison of unfolding transitions of single and double-substitutions suggests that replacement of the second (isolated) asparagine residue is modestly stabilizing, consistent with coupling of adjacent ladder asparagines through the hydrogen bonding network.

In addition to substitutions to asparagines 74 and 98 that disrupt the ladder, we attempted to extend the asparagine ladder of pp32 by introducing asparagine residues at equivalent ladder positions of repeats one, two, and five (serine 27, valine 52, and cysteine 123; Figure 1C). Asparagine substitutions in repeats one and two (S27N and V52N) were destabilizing (Figure S3), but substitution in repeat five (C123N) was stabilizing (Table 1). The C-terminal C123N variant has a far-UV CD spectrum similar to wild-type pp32 (Figure 2C). Although the increase in stability is significant, the magnitude of the $\Delta\Delta G^{\circ}_{H2O}$ value is less than that observed for hydrophobic substitutions of ladder asparagines 74 and 98 (Figure 2D, S2B and Table 1), suggesting that when in their native context, ladder asparagines are uniquely stabilizing.

The asymmetric effect of N- versus C-terminal extension on stability suggests local differences in context between repeats two and five. In attempting to determine the sequence origins of this variation, we identified a residue (threonine 49) in close proximity

to valine 52 that differs from the strongly conserved leucine normally found at this position (Figure 1C). Substitution of threonine 49 with the consensus leucine is highly stabilizing (Figure 2D, $\Delta\Delta G°_{H2O}$ = -2.6 kcal mol^{-1}); moreover, the far-UV CD spectrum of the T49L variant is nearly identical to wild-type pp32. The isosteric hydrophobic substitution T49V is also stabilizing, although the stability increment ($\Delta\Delta G°_{H2O}$ = -1.8 kcal mol^{-1}) is not as large as that for T49L (Table 1), suggesting that the large increase in stability of the T49L variant results both from substitution of the hydroxyl group with a methyl, as well as improved hydrophobic packing of the leucine side chain [29]. Although the T49L substitution significantly stabilizes the pp32 LRR domain, introduction of asparagine at position 52 is destabilizing in the T49L as well as the wild-type pp32 background (Table 1).

2.2.3 Assignments of backbone NH and asparagine side-chain NH_2 resonances.

Owing to the unique structural features of asparagine ladders, which involve burial and extensive intramolecular hydrogen bonding, the side chain NH_2 groups of ladder asparagines may be expected to have unique NMR signatures, including unusual chemical shifts and ^{1}H chemical shift temperature coefficients, which are sensitive to hydrogen bonding, as well as unique relaxation and exchange properties, which are sensitive to dynamics. Moreover, these structural features are likely to be sensitive to nearby perturbations such as ladder extension and changes in local stability and chemical environment. To explore these structural features, we assigned backbone resonances for the stabilizing T49L and C123N variants, as well as the side chain asparagine NH_2 resonances for wild type, T49L, and C123N pp32 variants.

The backbone of wild-type pp32 has been assigned previously using standard triple-resonance techniques [26]. Like wild-type pp32, the T49L, and C123N variants both display well-dispersed ^{1}H-^{15}N HSQC spectra (Figure S4A), with most resonances overlaying well with wild-type resonances. To confirm the identities of overlapping resonances and assign the resonances that differ from wild-type pp32, we collected HNCACB, CBCACONH, and HNCO spectra for T49L and C123N. From these spectra we were able to assign backbone resonances of 146 (T49L) and 151 (C123N) out of 152 residues.

To assign asparagine side-chain NH_2 resonances, we collected CBCGCO and NH_2-filtered versions of the ^{1}H-^{15}N HSQC and HNCO spectra to connect backbone assignments to side-chain ^{15}N nuclei. The NH_2-filtered ^{1}H-^{15}N HSQC experiment enhances resolution of asparagine and glutamine side chains by filtering out backbone amide groups via exploitation of the proton multiplicity of NH_2 versus NH groups (Figure 3A). Similarly, the NH_2-filtered HNCO experiment selectively correlate asparagine and glutamine side chain NH_2 protons with side chain CO nuclei. These spectra, in conjunction with CBCGCO spectra, were used to assign side chains starting with C_β assignments from HNCACB and CBCACONH experiments. Using this approach, we assigned all 14 asparagines in the T49L variant, and the 15 asparagines in the C123N variant (Figure S4B). The asparagine side chain assignments from C123N and T49L were readily transferred to wild-type pp32, YD, and L69A.

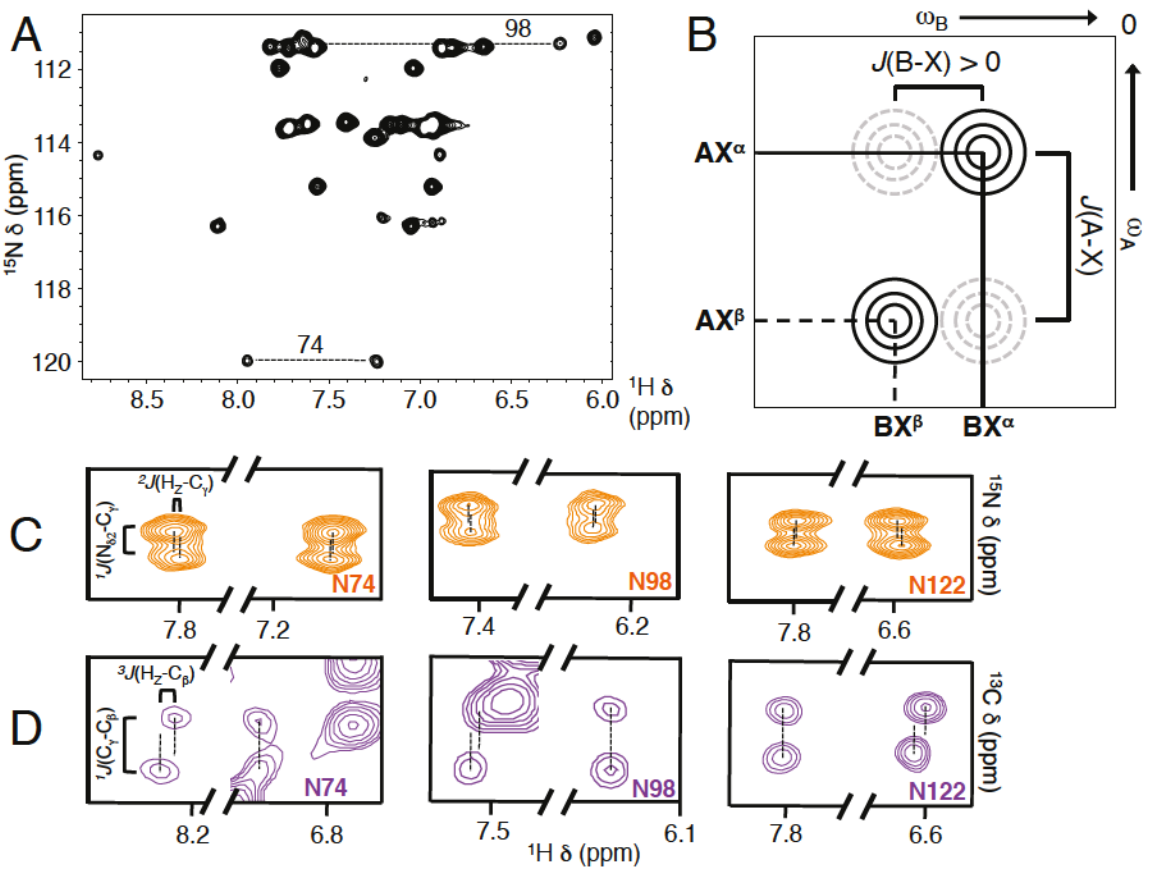


Figure 2.3. Stereospecific NMR assignment of asparagine NH_2 side chain protons (legend on following page).

NH_2 groups produce two resonances in ^{1}H-^{15}N HSQC spectra. For each pair, resonances have identical ^{15}N frequencies, but distinct ^{1}H frequencies (Figure 3A), corresponding to the H_Z and H_E protons (Figure 1A). For surface exposed asparagines, the H_Z proton is typically upfield of the H_E proton, since this proton is closer to the shielding $O_{\delta 1}$ electrons [30]. Because the unique structural environment of the ladder asparagines may produce significant chemical shift changes, we sought to directly assign the asparagine H_E and H_Z protons using E.COSY experiments [31], [32].

Figure 2.3. Stereospecific NMR assignment of asparagine NH_2 side chain protons. (A) Wild-type pp32 NH_2-filtered HSQC spectrum. Ladder side chain NH resonances are connected by dashed lines. For N98, only the upfield resonance is resolved, though the position of the downfield proton is known from HDX experiments where overlapping peaks exchange rapidly (Figure S7). (B) Schematic diagram of an E.COSY cross-peak pattern with a large positive *J*(A-X) coupling and a small positive *J*(B-X) coupling (brackets). Solid circles (black) represent observed resonances in which A/B nuclei are paired with the same X spin state (here, X^{α}) whereas dashed circles (grey) represent unobserved resonances in which A/B nuclei are paired with opposite X states (indicated by labels next to peaks). When *J*(A-X) and *J*(B-X) coupling constants have the same sign, the line connecting the observed resonances has a positive slope, as shown. *J*(A-X) and *J*(B-X) coupling constants of opposite sign have negative slope. (C) Selections from C123N $C_{\gamma}O$-coupled NH_2-HSQC E.COSY spectra to measure small $^{2}J(H_{E/Z}\text{-}C_{\gamma})$ values. $^{2}J(H_{E}\text{-}C_{\gamma})$ is positive (1 to 5 Hz, same as (B)) whereas $^{2}J(H_{Z}\text{-}C_{\gamma})$ is negative (-1 to -5 Hz), permitting the H_Z and H_E resonances to be assigned using the large positive $^{1}J(N_{\delta 2}\text{-}C_{\gamma})$ coupling constant. For the ladder asparagines 74 and 98, the upfield resonances are displaced with positive slope (as in B), identifying these resonances as originating from H_E. The downfield resonances are displaced with negative slope, identifying these resonances as originating from H_Z. For non-ladder asparagine residues such as N122 (right panel), this pattern is reversed. The full spectrum is shown in Figure S5. (D) Selections from T49L HNCO E.COSY to measure small $^{3}J(H_{E/Z}\text{-}C_{\beta})$ values. $^{3}J(H_{Z}\text{-}C_{\beta})$ is positive (5-10 Hz) whereas $^{3}J(H_{Z}\text{-}C_{\beta})$ is close to zero, permitting the H_Z and H_E resonances to be assigned using the large positive $^{1}J(C_{\gamma}\text{-}C_{\beta})$ coupling constant. For the ladder asparagines, downfield resonances are displaced with a positive slope, identifying these resonances as originating from H_Z, whereas upfield resonances have a vertical displacement, identifying these resonances as originating from H_E. Again, for non-ladder asparagine side chains (e.g., N122, right panel), this pattern is reversed. The full spectrum shown in Figure S5.

In E.COSY experiments a large *J*(A-X) coupling constant is exploited to accurately resolve a small *J*(B-X) coupling constant in a system of mutually *J*-coupled nuclei, A-X-B. Characteristic E.COSY peak patterns arise in A-B correlated 2D spectra in which the X nuclei are not decoupled during the experiment, ensuring cross peaks are only observed between A and B nuclei with matched X states (e.g., AX^{α}, BX^{α}, but not AX^{β}, BX^{α}, Figure 3B). We took advantage of relatively large negative $^{1}J(N_{\delta 2}\text{-}C_{\gamma})$ and large positive $^{1}J(C_{\gamma}\text{-}C_{\beta})$ to measure the small $^{2}J(H_{E/Z}\text{-}C_{\gamma})$ and $^{3}J(H_{E/Z}\text{-}C_{\beta})$ (Figure 3C and D respectively),

allowing us to determine stereospecific assignments for asparagine H_Z and H_E protons from the sign of the small $H_{E/Z}$-$C_{\beta,\gamma}$ coupling constants. The E.COSY experiments revealed that for asparagines 74 and 98, chemical shifts of H_E and H_Z protons are inverted with respect to the pattern expected for surface asparagine residues, with the H_Z protons downfield relative to the H_E protons (Figure 3C, D; Figure S5). These assignments are consistent with a number of NOEs between the asparagines 74 and 98 NH_2 protons and protons that are nearby in the crystal structure of pp32 (data not shown). The inversion of the H_E and H_Z chemical shifts is unique to the asparagine ladder residues, as all of the 12 other asparagine side chains have H_E chemical shifts downfield of their H_Z chemical shifts, including asparagine 123 in C123N (Figure 4A). This inversion is unlikely to be due to local magnetic fields from nearby aromatics as the nearest aromatic group is more than 10 Å from either asparagine ladder NH_2 group.

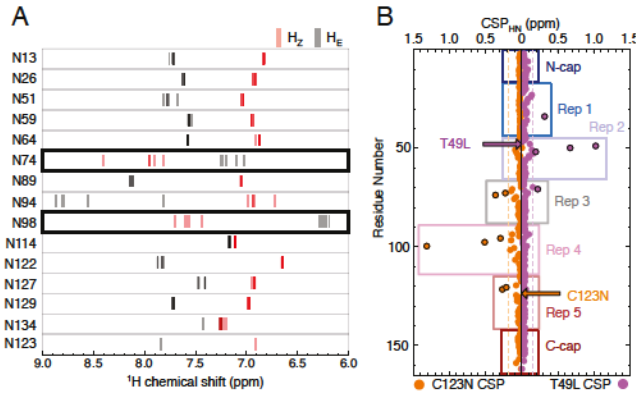


Figure 2.4. Asparagine side chain proton chemical shifts and chemical shift perturbations (CSPs) of backbone amides in pp32 variants relative to WT pp32. (A) H_E and H_Z chemical shifts for each asparagine side chain from wild-type pp32 and peripheral variants are shown as gray and red lines, respectively. Ladder asparagines 74 and 98 are outlined in black. (B) CSPs for T49L (purple) and C123N (orange) backbone amides. Black outlined circles denote perturbations ≥ 1 standard deviation (σ_{CSP}) from the mean (dashed vertical lines). Boxes indicate repeat and cap boundaries. Arrows indicate the location of each substitution.

Hydrogen bonding usually results in large downfield movements of proton chemical shifts [33]. Measurement of the chemical shifts of asparagines 74 and 98 in wild-type pp32 provides a qualitative estimate of the strength of hydrogen bonding in these ladder residues. The inversion of H_Z and H_E proton chemical shifts results from movement of H_Z resonances downfield by 0.6 to 1 ppm and H_E resonances upfield by 0.4 to1.3 ppm (Figure 4A). This inversion suggests strong H_Z hydrogen bonding with the *i-27* backbone carbonyl group of the previous repeat for both asparagines and weak H_E hydrogen bonding to the *i-3* CO group. Note that the non-ladder asparagine 94 H_E proton is also shifted quite far downfield in all sequence backgrounds, consistent with strong hydrogen bonding. In the high-resolution crystal structure of pp32 (PDB ID: 4XOS), the H_E of asparagine 94 makes a 2.8 Å hydrogen bond with a side-chain carboxylate oxygen from aspartate 70.

2.2.4 Chemical shift sensitivities of backbone NH resonances to N- and C-terminal structural perturbation.

Asparagine ladders provide a direct network of hydrogen bonding from the N-terminus to the C-terminus of LRR proteins and may propagate local changes over large distances. To examine whether such propagation occurs in pp32, we examined the effects of two peripheral mutations (T49L and C123N) on the chemical shifts of nearby and distant residues. These variants were chosen because both are stabilizing and both maintain structured native states, based on CD and NMR spectroscopy, and because together they provide a comparison of perturbations from a potential ladder-extending variant (C123N) with a non-extending variant (T49L).

Comparison of backbone amide proton chemical shifts of the T49L and C123N variants to those of wild-type pp32 shows that the C123N substitution causes more distant ^{1}H-^{15}N chemical shift perturbations (CSPs) than the T49L variant (Figure 4B). In the C123N variant, backbone CSPs extend N-terminally from the site of the substitution. The largest backbone CSPs in C123N are in repeat four, adjacent to the substituted repeat. Though the backbone CSPs in repeat three are smaller than in repeat four, they are of the same magnitude as those in repeat five, the site of the C123N substitution. The ladder asparagines 98 and 74 have the second and third largest backbone CSPs respectively (Figure S6A). Propagation of structural changes from the C-terminal end of the asparagine ladder (98) to the N-terminal end (74) may be responsible for the large CSPs observed far from the substitution site. In contrast, CSPs of the T49L variant are more localized to the site of substitution and do not affect the backbone of either asparagine ladder residue (Figure S6B). Since both the T49L and C123N substitutions are stabilizing, the longer-range perturbations of the C123N substitution, especially at ladder residues 74 and 98, may reflect propagation of perturbations through the ladder.

Given the importance of side chain interactions in the asparagine ladder, CSPs resulting from T49L and C123N substitution were also determined for the $H_{E/Z}$ protons of all 14 asparagines (Figure S6C). In T49L, ladder side chain CSPs extend C-terminally, decreasing in magnitude with increasing distance from the substitution site (that is, N74 H_Z >> N74 H_E > N98 H_Z; the chemical shift of the N98 H_E is not significantly perturbed). These side-chain CSPs extend a full repeat further than backbone CSPs for T49L. In C123N, however, ladder side chain CSPs are significant only for H_Z protons, even though

asparagine 98 H_E is closest to the substitution site. Thus, for both N- and C-terminal substitution (T49L and C123N), the H_Z protons are more sensitive to perturbation than the H_E protons, which may be related to their large downfield chemical shifts, and perhaps stronger hydrogen bonding, compared to the H_E protons.

2.2.5 Dynamics of asparagine 74 and asparagine 98 side-chain NH_2 groups.

One conspicuous feature of the NH_2 cross peaks of the asparagine ladder residues is their low intensities compared to surface asparagine cross peaks (Figure 3A). This decreased intensity is seen in all peripheral variants, though interestingly, the NH_2 resonances of potential ladder extending asparagine 123 are of high intensity, comparable to solvent-exposed asparagines. Reduced intensity could either result from large amplitude dynamics on the chemical shift timescale (μs-ms), or from rapid transverse relaxation (large R_2 values) due to both the slow overall tumbling on the nanosecond timescale and a higher local concentration of hydrogen spin density, compared to surface asparagine NH_2 groups. To resolve these two possibilities and to investigate the overall rigidity of the asparagine ladder, we measured ^{15}N spin relaxation of the asparagine side chain NH_2 groups in pp32 using various NMR experiments that probe dynamics on different timescales.

Due to the low signal intensities of asparagine 74 and 98 side chain NH_2s, conventional relaxation experiments suffered from low signal-to-noise. To increase signal intensity, we equilibrated ^{15}N-labelled proteins in a 7 M urea solution containing 50% D_2O. Under these conditions, all pp32 variants are unfolded, and undergo rapid hydrogen exchange with solvent. Upon refolding, exchangeable sites should be 50% deuterated on

average. For the asparagine side-chain amides, this level of deuteration leads to an equal distribution of the four isotopomers (NH_2, NH_ZD_E, NH_ED_Z, and ND_2). The singly-deuterated species have significantly reduced H_Z-H_E dipole-dipole relaxation compared to the NH_2 species. Combining an NH-filtered pulse sequence with deuterium-decoupling to reduce scalar relaxation of the second kind suppresses signals from the NH_2 species. Thus, the only cross peaks in this experiment are those of the singly protonated species (NH_ZD_E and NH_ED_Z), which have higher signal-to-noise due to elimination of H_Z-H_E dipole-dipole relaxation. Thus, standard pulse sequences can be used to probe ^{15}N dynamics (with minimal modifications such as ^{2}H decoupling) while retaining high signal-to-noise.

To test whether the reduced signal intensities of asparagines 74 and 98 result from intermediate exchange on the μs-ms timescale we conducted "two-point" relaxation dispersion (RD) constant time (CT) CPMG experiments [34]. In these experiments, two spectra are acquired with CPMG spin echoes applied with different spacings over a constant time period (T), typically 35 to 40 ms in duration. During this period, the NMR signal decays according to an apparent relaxation rate $R_2^{app} = R_2 + R_{ex}(\nu)$ where R_2 is the contribution from ps-ns timescale dynamics and $R_{ex}(\nu)$ is the ms-μs contribution, which depends on the time interval between the spin echo pulses. The peak intensity at the end of the CT period (T) is given by $I(T) = I_0 e^{-R_2^{app} T}$. In the first spectrum (A) the frequency of spin echoes is kept at the maximum value allowed by duty cycle considerations, minimizing the effects of exchange broadening on signal decay. If exchange contributions are completely quenched, $R_2^{app} = R_2$. In reality, $R_2^{app} = R_2^* \sim R_2$ where R_2^* is a best-case approximation to the ps-ns timescale dynamics contribution to the signal decay.

Therefore, $I(A) = I_0 e^{-R_2^* T} \sim I_0 e^{-R_2 T}$. In the second spectrum (B) the frequency of spin echoes is minimized, thereby maximizing signal decay from exchange broadenings. In this scenario, $R_2^{app} \sim R_2 + R_{ex}$ where R_{ex} is the ms-μs contribution to the signal decay. As a result, $I(B) = I_0 e^{-(R_2 + R_{ex})T}$. From this it follows that cross-peak intensities with significant R_{ex} contributions will have lower peak intensities in (B) than in (A). Typically, I(B)/I(A) ratios < 0.75 are considered to be "candidates" for μs-ms dynamics. Additionally, spectrum A can be used to calculate $R_2^* \sim R_2$, the exchange-minimized transverse relaxation rate, which is sensitive to ps-ns timescale rotational diffusion and N-H bond vector librational dynamics.

The two-point experiments did not show any difference in signal intensity between the A and B spectra for asparagine ladder residues in any variant, indicating that the asparagine ladder is rigid on the μs-ms timescale. Thus, fast to intermediate exchange dynamics of ^{15}N nuclei are unlikely to cause the decreased intensities observed for ladder NH_2 resonances. Rather, the CPMG experiments reveal that asparagine 74 and 98 side chain NH(D) R_2^* values are among the largest relaxation rates of the asparagine side-chains in pp32 (Figure 5A). In fact, asparagines 74 and 98 side chains have R_2^* values equal to or greater than most backbone R_2^* values (Figure 5A, compare black outlined points to dashes). This is true for all variants at both 20 and 35 °C, suggesting that the side chains of asparagines 74 and 98 have rigidity (and hence, rotational and correlation times) comparable to hydrogen-bonded backbone NH groups. In contrast, R_2^* values for the N123 side chain NH(D) groups in the C123N variant are much lower (both at 20 and 35°C) than for the 74 and 98 side chains.

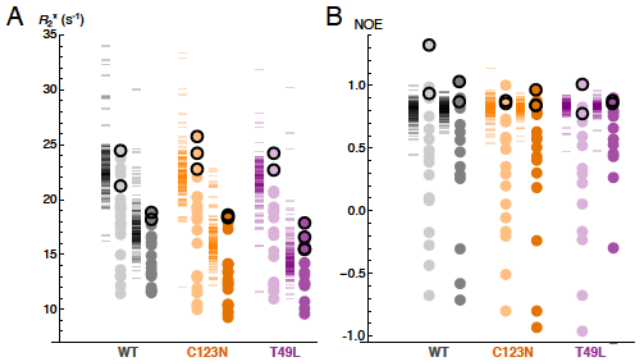


Figure 2.5. Transverse relaxation rates and ^{1}H-^{15}N NOEs for asparagine NHD and backbone NH groups in wild-type pp32 and stabilizing variants. Circles indicate asparagine side-chain NHD values, dashes to the left of circles indicate backbone amide values. Experiments were performed at 20 °C and 35 °C (light and dark colors, respectively). Values from N74 and N98 side chains are outlined in black. Samples were ^{15}N-labelled and fully exchanged into 50% D_2O to improve signal-to-noise of ladder side chain resonances. (A) R_2^* measurements in wild-type pp32, T49L, and C123N. R_2^* measurements were determined from two-point CPMG experiments [34] with $R_2^* = -T^{-1}ln(I_{max}/I_0)$ where T is the constant CPMG time period used for the experiment, I_{max} is peak intensity with ν_{cpmg} set at the maximum value during the T period, and I_0 as the reference peak intensity. (B) ^{1}H-^{15}N NOE values for wild-type pp32, T49L, and C123N. Although the NOE value for the wild-type asparagine 98 H_Z appears to exceed the theoretical maximum of ~0.85 at 20 °C [35], its value at 35 °C is within the expected range for a protein of this size. Conditions: 20 mM $NaPO_4$, 50 mM NaCl, and 0.1 mM TCEP, pH 6.8 (after accounting for D_2O).

To explore rigidity and dynamics on a faster timescale, we measured ^{1}H-^{15}N NOEs for wild-type pp32 and the T49L and C123N variants, again using 50% deuterated samples. All ladder asparagines NH(D) groups have ^{1}H-^{15}N NOE values of approximately one. In contrast, non-ladder asparagine NH(D)s span a broad range of ^{1}H-^{15}N NOE values with an average of 0.29. As with R_2^* values, the ^{1}H-^{15}N NOE values of asparagine 74 and 98 protons are similar to values for rigid backbone amides (Figure 5B, compare black outlined points to dashes).

2.2.6 Temperature coefficients of asparagine 74 and asparagine 98 side chain NH_2 groups.

The dynamics experiments above confirm the ladder is rigidly structured, consistent with formation of strong hydrogen bonds. Strong hydrogen bonding involving the asparagine ladder H_Z protons is also suggested from the chemical shift values. To further probe the hydrogen bond strength of the asparagine ladder side chains, we measured the magnitude of the change in proton chemical shifts with increasing temperature (so-called "temperature coefficients", $\Delta\delta_{NH}/\Delta T$) for all asparagine side chain NH_2 groups for wild-type pp32, T49L, L69A, C123N, and YD (Figure 6). $\Delta\delta_{NH}/\Delta T$ values can be used to identify intramolecular hydrogen bonding [36] and have been shown to correlate with hydrogen bond length [37]–[39] and local unfolding [40]. Amides engaged in intramolecular hydrogen bonds have $\Delta\delta_{NH}/\Delta T$ values greater (i.e., less negative) than -4.5 ppb K^{-1} [41]; within this range, short strong hydrogen bonds tend to have more negative $\Delta\delta_{NH}/\Delta T$ values [39].

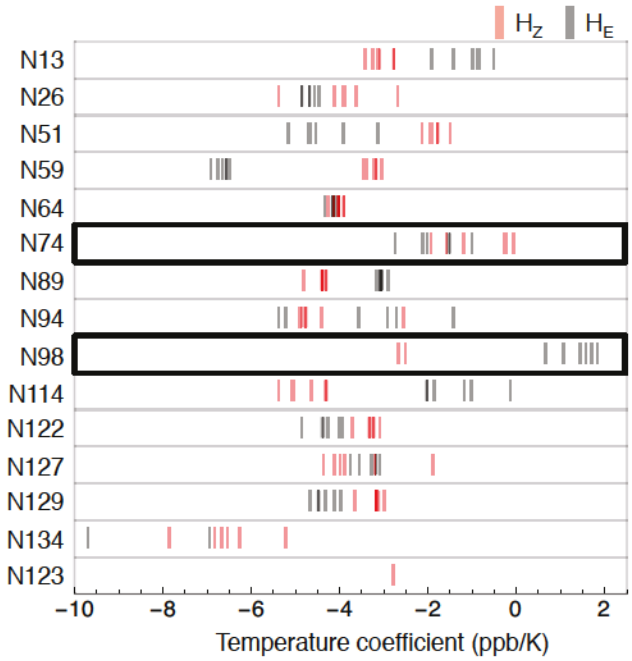


Figure 2.6. Temperature coefficients for asparagine side chains in wild-type pp32 and peripheral variants. H_E and H_Z temperature coefficients for each asparagine side chain from wild-type pp32 and peripheral variants are shown as gray and red lines, respectively. The two ladder positions are outlined in black. Temperature coefficients were determined from linear fits to changes in proton chemical shift at four temperatures (283, 288, 293, 303 K). Conditions: ~600 μM protein, 20 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, pH 6.8.

The non-ladder asparagine $\Delta\delta_{NH}/\Delta T$ values are all negative, with mean values of -3.5 ± 1.4 ppb K^{-1} (Figure 6). For some but not all of these non-ladder asparagines, H_Z and H_E $\Delta\delta_{NH}/\Delta T$ values are distinct and are largely independent of the variant background (for example, N59 and N114). The ladder asparagines have more positive $\Delta\delta_{NH}/\Delta T$ values than most non-ladder asparagines (Figure 6). This indicates greater resilience to temperature perturbation, as expected for stable intramolecular hydrogen bonds. Asparagine 74 H_Z $\Delta\delta_{NH}/\Delta T$ values are generally more positive than those of H_E; the sole exception is that of the T49L variant. In contrast, asparagine 98 H_E $\Delta\delta_{NH}/\Delta T$ values are

larger than those of H_Z, with the former showing positive $\Delta\delta_{NH}/\Delta T$ values. This positive $\Delta\delta_{NH}/\Delta T$ value suggests that the hydrogen bond involving asparagine 98 H_E is weak, consistent with its measured ^1H chemical shift. In contrast, the asparagine 98 H_Zs (measured only for C123N and T49L variants) are among the most negative ladder $\Delta\delta_{NH}/\Delta T$ values, suggesting that the hydrogen bond involving asparagine 98 H_Z is unusually strong, consistent with its measured ^1H chemical shift.

2.2.7 Hydrogen exchange of asparagine ladder NH_2 groups.

Chemical shift values, dynamics, and temperature coefficients provide convincing evidence for the existence of a rigid network of bonds connecting the side-chains of asparagines 74 and 98, but they do not provide estimates for the stability of the asparagine ladder structure. The rates of hydrogen exchange of amide groups with solvent are influenced by the stability of hydrogen-bonds involving amide NH groups, since exchange requires disruption of intramolecular hydrogen bonding [42]. Thus, NH groups that are strongly hydrogen bonded exchange slowly. For particularly stable hydrogen bonds, exchange rates may decrease to the limit set by global stability. Although global protection from exchange is often observed for a subset of backbone NH groups that are buried and stably hydrogen bonded into secondary structures, large protection factors are not typically seen for labile side-chain protons [43].

To probe the local stability of the asparagine ladder, we monitored hydrogen-deuterium exchange rates (HDX) by rapidly changing the solvent from H_2O to D_2O, and collecting ^1H-^{15}N HSQC spectra over time (Figure 7). The NH_2-filtered HSQC pulse sequence was used to reduce spectral overlap between asparagine 74 and 98 NH_2

groups and backbone amides. As a result, exchange curves report on the decay of the NH_2 isotopomer to either the NH_ED_Z or the ND_EH_Z isotopomer. Therefore, the fitted exchange rate constant is the sum of the individual H_E and H_Z exchange rate constants, and can be regarded as an upper limit for exchange. If one hydrogen exchanges faster than the other, the rate constant approximates the fast exchange process, whereas if the two protons exchange at the same rate, the fitted rate constant is twice the individual exchange rate constants.

In the first spectrum obtained after exchange is initiated (25 to 90 minutes), all of the solvent-exposed asparagines have fully exchanged, leaving only cross peaks for N74 and N98 NH_2 groups (Figure S7, note that rapid exchange during sample preparation is also observed for the potential ladder extending asparagine 123 in the C123N variant). The side-chain NH_2 protons of asparagine 74 are highly protected, with lifetimes ranging from hours to days depending on the variant (Figure 7A). The side-chain NH_2 protons of asparagine 98 are even more protected than those of asparagine 74, with lifetimes up to two months for the most stable variants (Figure 7B). The exchange rates of asparagine 74 and 98 are several orders of magnitude smaller than those measured for solvent exposed asparagines [43].

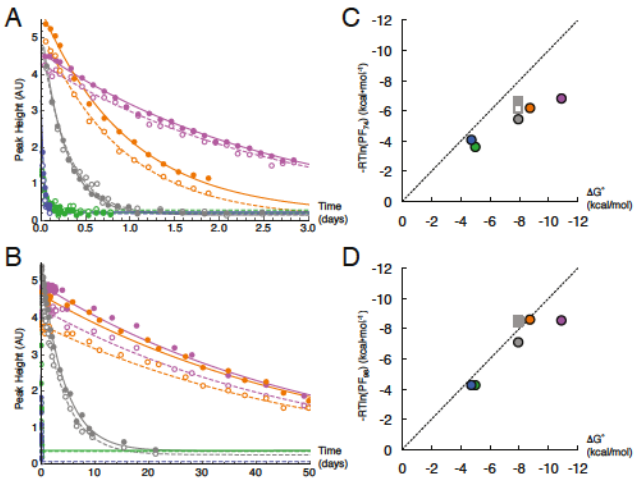


Figure 2.7. Side-chain hydrogen exchange data and protection factors for N74 and N98 side-chain NH_2 groups in wild-type pp32 and variants. (A) N74 peak heights as a function of exchange time for wild-type pp32 and peripheral variants (wild-type, gray; T49L, purple; C123N, orange; L69A, green; and YD, blue). H_E heights are shown as unfilled circles with dashed fitted curves. H_Z heights are shown as filled circles with solid fitted curves. (B) N98 peak heights as a function of exchange time as in (A). (C) Logarithmic protection factors for N74 NH_2 protons versus folding free energies calculated from urea denaturation experiments (Figure 2). Circles are from NH_2-filtered HSQCs are colored as in (A); squares are from unfiltered HSQC spectra of wild-type pp32 [21] (H_Z, filled gray square; H_E, empty gray squares). The dashed line represents global exchange (slope = 1, intercept = 0). (D) Logarithmic protection factors for N98 NH_2 protons as in (B). Conditions: 150 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, pH 6.8 (after accounting for D_2O) at 30 °C. Note that although these conditions are slightly different than those in Figure 2D, these differences have no effect on fitted thermodynamic parameters (not shown).

Though the individual H_E and H_Z exchange rates could not be determined from the NH_2-filtered exchange data, we were able to determine H_E and H_Z exchange rates for both ladder asparagines in wild-type pp32 using unfiltered ^{1}H-^{15}N-HSQC spectra that were obtained in a previous study of backbone amide hydrogen exchange [21]. Using these spectra, we determined the H_E and H_Z exchange rates for the side chains of asparagine 74 and 98 by monitoring cross peaks from the NHD species (Figure S8). The asparagine 74 H_Z exchanges more slowly than the H_E (Table 2), consistent with H_Z and

H_E exchange rates in model compounds [44] and with an EX2 exchange process. The asparagine 98 H_E and H_Z exchange at nearly the same rate, which may be an indication that for this highly protected group, exchange has some EX1 character. The overall slower rates of exchange determined from the unfiltered versus filtered HSQC spectra is likely to result from the lower temperature of the former (20 °C versus 30 °C).

Table 2. Hydrogen exchange rates for ladder asparagine side chains in pp32 variants.

	Asparagine 74			Asparagine 98		
Variant	NH_2[a]	H_Z[b]	H_E[b]	NH_2[a]	H_Z[b]	H_E[b]
wild-type	$41.7 \pm 1.7 \times 10^{-6}$	$3.0 \pm 0.3 \times 10^{-6}$	$7.7 \pm 1.3 \times 10^{-6}$	$2.7 \pm 0.1 \times 10^{-6}$	$1.4 \pm 0.2 \times 10^{-7}$	$1.2 \pm 0.2 \times 10^{-7}$
T49L	$4.3 \pm 0.1 \times 10^{-6}$	–	–	$2.3 \pm 0.1 \times 10^{-7}$	–	–
L69A[c]	$845 \pm 92 \times 10^{-6}$	–	–	$301 \pm 58 \times 10^{-6}$	–	–
C123N	$12.4 \pm 0.2 \times 10^{-6}$	–	–	$2.3 \pm 0.1 \times 10^{-7}$	–	–
YD[c]	$38 \pm 9 \times 10^{-6}$	–	–	$17 \pm 8 \times 10^{-6}$	–	–

Units for hydrogen exchange rate constants are s^{-1}. Uncertainties estimates are from the square-root of the diagonal elements of the covariance matrix obtained from nonlinear least-squares fitting. [a]Hydrogen exchange rates were measured from NH_2 filtered HSQC spectra at 30 °C with 150 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, pH 6.8 (after correcting for D_2O). [b]Hydrogen exchange rates were calculated using unfiltered HSQC spectra from [21] collected at 20 °C with 20 mM $NaPO_4$, 50 mM NaCl, 0.1 mM EDTA, 0.2 mM TCEP, pH 6.7 (after correcting for D_2O). [c]Only partial decay curves were obtained due to rapid exchange; thus, uncertainties in rate constants are comparatively large.

To determine whether exchange of the asparagine ladder NH_2 protons requires complete unfolding for exchange, we computed local stabilities from protection factors (PFs) and compared these values to unfolding free energies measured from urea denaturation experiments. The ladder extending (C123N) and peripheral substitutions (T49L, L69A, YD) were used to modulate global stability, and potentially alter hydrogen exchange rates. For asparagine 74, local stabilities estimated from hydrogen exchange measurements are lower than the global stability limit for each variant, particularly for variants with increased global stability (C123N and T49L; Figure 7C). This suggests that exchange of the N74 side-chain NH_2s involves a sub-global mechanism, although the

partial correlation to global unfolding free energies indicates that the exchange-competent forms are influenced to some degree by overall stability. In contrast, local stabilities estimated from asparagine 98 hydrogen exchange rates are close to values expected from global exchange (Figure 7D) over a broad range of global stabilities. Only the most highly stable T49L variant falls off the unit slope line, exchanging with a rate that is roughly equivalent to that of the C123N variant despite being 1.6 kcal mol^{-1} more stable. These very slow hydrogen exchange rates are consistent with strong hydrogen bonding for both ladder asparagine sidechains, especially that of asparagine 98.

2.3 Discussion

The asparagine ladder is a highly conserved feature of many LRR proteins. Despite conjecture about the importance of the ladder to LRR proteins [23], few studies have directly probed the role of the asparagine ladder in LRR protein structure and stability. The unique properties shared by asparagine 74 and 98 (i.e. inverted H_E and H_Z chemical shift, large protection factors, large contributions to $\Delta G°_{H2O}$, etc.) derive from their chemical environment, which crystal structures show to be common among ladder asparagines across the LRR family [22], [23], [45]. Thus, it is likely that the features observed in the short asparagine ladder of pp32 are representative of residues in longer ladders like those of YopM. The present study shows the asparagine ladder is a key component of LRR structure that forms a rigid network of hydrogen bonds providing repeat-to-repeat coupling.

2.3.1 Asparagine ladder hydrogen bonds.

Crystal structures have shown that asparagine ladder side chains form hydrogen bonds with local backbone atoms [23]. The urea-induced unfolding transitions measured here provide compelling experimental evidence that these hydrogen bonds contribute favorably to stability (Figure 2). Substitution of either ladder residue increases the folding free energy of pp32 by about 4.5 kcal mol^{-1} compared to wild-type pp32, Table 1. Ladder asparagine side chains also behave like structured backbone amides in NMR relaxation experiments (Figure 5). Additionally, ladder side chains are highly protected from hydrogen exchange (Figure 6). To our knowledge, the only comparable level of protection of exchangeable side chains are asparagines 43 and 44 of BPTI [46]. However, unlike the ladder asparagine 98 in pp32, the protected asparagines in BPTI exchange more rapidly than the global exchange limit, based on the stability of BPTI [47]. Given these data, it is clear that ladder side chains form stable interactions integral to the LRR motif.

The high degree of protection of asparagine ladder NH_2 groups from hydrogen exchange demonstrates that these groups are completely sequestered from solvent. Thus, variations in the temperature coefficients ($\Delta\delta_{NH}/\Delta T$ values) of ladder NH_2 proton resonances result from changes in the native protein structure. The asparagine 98 H_E and H_Z $\Delta\delta_{NH}/\Delta T$ values are particularly notable given their opposite signs compared to surface exposed asparagines, with H_E protons showing positive $\Delta\delta_{NH}/\Delta T$ values (Figure 6). The large difference between in $\Delta\delta_{NH}/\Delta T$ values for the H_E and H_Z protons of asparagine 98 suggests different levels of hydrogen bonding for the H_E and H_Z proton.

As with $\Delta\delta_{NH}/\Delta T$ values, the H_E and H_Z protons of asparagine 98 also differ significantly in their proton chemical shifts, which are not only inverted compared to

unstructured asparagine NH_2 groups but are significantly separated from each other (by about 1.35 ppm in the proton dimension, Figure 4A). The downfield shifts in the H_Z asparagine 98 protons and the upfield shifts of the H_E protons relative to surface asparagines suggests an asymmetric degree of hydrogen bonding by the ladder side-chains, in which the H_Z hydrogen forms a strong hydrogen bond to the previous repeat, whereas the H_E proton hydrogen bonds weakly. This is supported by the $\Delta\delta_{NH}/\Delta T$ values, which shows that the asparagine 98 H_Z is more negative than that of H_E consistent with a shorter H_Z hydrogen bond [38].

Though the H_E/H_Z chemical shift differences for asparagine 74 are similar to those for asparagine 98, they are less pronounced, consistent with the overlapping distribution of $\Delta\delta_{NH}/\Delta T$ values for the H_Z and H_E resonances of asparagine 74. Interestingly, the T49L variant is exceptional in this regard, showing an H_Z/H_E proton chemical shift separation (1.380 ppm) that is similar to those for asparagine 98 (~1.48 ppm); likewise, the asparagine 74 H_E $\Delta\delta_{NH}/\Delta T$ value in the T49L variant is larger than that of H_Z. It is possible that the strongly stabilizing T49L substitution strengthens the asparagine 74 H_Z hydrogen bond to the backbone CO group of residue 49 (the site of the substitution) so that it resembles that of asparagine 98.

The unique spectroscopic features of asparagine 98 (in particular, chemical shifts, $\Delta\delta_{NH}/\Delta T$ values), and the convergence of asparagine 74 to these features in a variant in which the donor residue to the asparagine 74 H_Z is substituted to the consensus sequence (Figure S1) suggest an archetypal LRR ladder bonding pattern in which the inter-repeat H_Z hydrogen bond is strong, and the intra-repeat H_E hydrogen bond is weak.

This pattern of bonding is consistent with the absence of a detectable isotope effect between asparagine 74 H_E and asparagine 98 H_Z despite their sharing a hydrogen bond acceptor (Figure 1A). Such isotope effects have been observed between hydrogen bond pairs to a single acceptor [48], [49]; the absence of such an effect here, along with the upfield chemical shift of H_E, is consistent with a weak H_E hydrogen bond.

2.3.2 Asparagine ladder structural features.

Above, we described the evidence for the asparagine ladder's formation of stable hydrogen bonds. As a result, the highly structured side chains likely experience increased dipolar relaxation from other nearby structured protons, which would explain why the ladder $H_{E/Z}$ protons have low signal intensity. High-resolution structures of pp32 show that asparagine 74 and 98 $H_{E/Z}$ are indeed surrounded by a larger number of ordered protons than non-ladder asparagines (Figure S9A). In addition, backbone amide signal intensity is inversely correlated with the number of interproton contacts (Figure S9B). Since the ladder side chains seem to behave like the structured backbone amides, we view the asparagine ladder as a "second backbone", providing a hydrogen bonding network that extends through the hydrophobic core. In large LRR proteins such as YopM, extended asparagine ladders may provide coupling over long distances, and may contribute to the high degree of cooperativity seen in those proteins [20], [50].

The importance of this second backbone to the LRR motif is exemplified by the far-UV CD spectra of wild-type pp32 and ladder substituting variants (Figure 2A). These variants show progressively larger disruptions in native secondary structure as one and then both asparagines are substituted. YopM, another LRR protein, shows similar

changes in its far-UV CD spectrum after deletion of stabilizing repeats, which results from unfolding of multiple repeats that are adjacent to the site of deletion [51]. These partial unfolding transitions highlight the importance of interfaces in stabilizing leucine-rich repeats. The pp32 data extends this observation, showing that only the conserved asparagine need be removed to elicit a similar disruption in secondary structure. This suggests that the asparagine ladder is a major contributor to LRR interfaces, which are integral to LRR secondary structure.

Despite the demonstrated importance of the asparagine ladder to LRR structure, ladder-extending substitutions were either destabilizing (S27N, V52N; Figure S3) or only marginally stabilizing (C123N, Figure 2D). For C123N, the substituted asparagine has small R_2^*, undergoes rapid hydrogen exchange (Figures 5A and 6), is highly dynamic on the ps-ns timescale (Figure 5B), and lacks strong NOEs to protons expected to be in close proximity (data not shown), features more similar to the solvent-exposed asparagines than to ladder asparagines 74 and 98. The inability to extend the ladder to adjacent repeats highlights the importance of sequence context for the ladder architecture. It appears that LRRs lacking an asparagine at the ladder position have additional sequence differences that stabilize the non-asparagine residue (e.g., valine 52 in pp32), and that these sequence differences are incompatible with an asparagine. However, multiple sequence alignments of tandem LRRs show little pairwise covariance between the ladder position and other positions, suggesting that any such covariance involves multiple positions.

2.3.3 The asparagine ladder and cooperativity.

As a result of the asparagine ladder's role at repeat interfaces, it is likely to contribute to cooperativity during protein folding. The chemical denaturation of the N98A variant of pp32 supports this hypothesis, showing a significantly shallower unfolding transition than wild-type pp32 while retaining native-like structure. This behavior differs from that of pp32 variants that replace the conserved leucine residues that define the LRR motif. Previous studies have shown that substitutions to conserved leucines are similarly destabilizing but do not affect cooperativity [20], [21]. Coupling is also demonstrated in the non-additivity of ΔG°_{H2O} values for the asparagine ladder substitutions. Substituting either ladder asparagine is strongly destabilizing ($\Delta\Delta G^{\circ}_{H2O} \approx$ +3.5 kcal mol^{-1}), but substitution of the second ladder asparagine is modestly stabilizing ($\Delta\Delta G^{\circ}_{H2O} \approx$ -0.4 kcal mol^{-1} relative to the single variant). This indicates an energetic coupling between the two positions [52], [53] of around -5 kcal mol^{-1}; in other words, ladder asparagines are stabilized by their neighboring asparagines by 5 kcal mol^{-1}. Since individual LRRs are likely unfolded [54], [55], the coupling provided by the asparagine ladder is important for LRR protein folding.

Although the double mutant cycles reveal coupling between ladder asparagines, they do not reveal how coupling is achieved. A possible mechanism is suggested from the pattern of CSPs in the stabilizing pp32 variant C123N. This substitution produces significant CSPs for asparagine 74 backbone NH and side chain H_Z despite their distance from the substitution site (> 9 Å). Since asparagine 98 directly contacts the substituted residue 123, residue 74 CSPs may be transmitted through structural rearrangements of asparagine 98. This would imply the asparagine ladder residues can propagate structural

changes between adjacent ladder positions. In addition to the hydrogen bonds between the asparagine side chain NH_2 and backbone carbonyl oxygens at positions *i-27* and *i-3*, a second possible conduit for propagation is hydrogen bonding between the backbone amide of the ladder residue in repeat *i-1* and the ladder side chain $O_{\delta 1}$ in repeat *i* (Figure 1A). Perhaps the modest sensitivity of the H_E chemical shift to substitution is related to the comparatively weak hydrogen bond involving the H_E proton. Small structural changes would result in large chemical shift changes for the strongly hydrogen bonded H_Z (consistent with large negative $\Delta\delta_{NH}/\Delta T$ values) but small chemical shift changes for the weakly bonded H_E proton. This explanation is consistent with the observation that the H_Z CSP resulting from C123N substitution is in the upfield direction (Figure S6).

2.4 Materials and Methods

2.4.1 Protein Cloning, Expression, and Purification.

Protein expression and purification were performed as described in [26]. Mutated versions of the pp32 gene were generated using the Quikchange mutagenesis kit (Agilent Technologies). For ^{15}N (and ^{13}C) NMR samples, bacteria were grown in M9 minimal media supplemented with $^{15}NH_4Cl$ (and ^{13}C glucose). Protein preparation was performed as described in [26]. NMR samples were dialyzed into 20 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, pH 6.8 prior to data acquisition.

For side chain dynamics and E.COSY experiments, labeled protein was partially deuterated by unfolding in 50% ^2H denaturant buffer (8 M Urea, 50 mM NaCl, 0.1 mM

TCEP, 20 mM $NaPO_4$ pH 6.8 (after adjustment for deuterium)) for at least 30 minutes to allow for complete exchange. 50% ^2H refolding buffer (50 mM NaCl, 0.1 mM TCEP, 20 mM $NaPO_4$ pH 6.8 (after adjustment for deuterium)) was then added to denatured samples to refold the protein and samples were concentrated using filtration columns (GE Healthcare). Residual denaturant was removed with a spin column [56].

2.4.2 Circular dichroism spectra and equilibrium unfolding.

All CD experiments were performed on an Aviv Model 400 CD spectropolarimeter using a computer-controlled Hamilton Microlab syringe titrator with samples in CD buffer (20 mM NaPO4, 150 mM NaCl, 0.1 mM TCEP, pH 7.8) at 20 °C. CD spectra were collected with 5 second signal averaging every nm from 280 to 200 nm; protein concentrations were 30 to 50 μM in a 0.1 cm path-length quartz cuvette. Equilibrium unfolding experiments were monitored at 220 nm using a 5-minute equilibration time and 30 second signal averaging; protein concentrations were 3-5 μM in a 1 cm path length quartz cuvette. Urea (VWR Life Sciences) used for denaturation studies was deionized using a mixed-bed resin (BioRad) immediately prior to use; urea concentrations were determined using refractometry [57]. Two-state analysis of equilibrium unfolding experiments was performed as described in [58]. Errors in thermodynamic parameters from equilibrium unfolding experiments are from three independent experiments.

2.4.3 NMR spectroscopy.

Backbone resonances of wild-type and variant pp32 proteins were assigned as in [26] using a Bruker Avance II 600 MHz spectrometer equipped with a cryoprobe. Triple-resonance experiments (HNCA, HNCACB, CBCA(CO)NH, HNCO, ^{15}N-edited ^1H-^1H

NOESY) along with wild-type pp32 assignments [26] were used to assign variant proteins. Assignments were made using the CARA program [59].

Side-chain assignments were determined using NH_2-filtered HSQC, CBCGCO [60], and C□O-coupled NH_2-HSQC/HNCO E.COSY experiments [31], [32], [61]. Side chain C_γ assignments were made using CBCGCO spectra in conjunction with C_β assignments from backbone triple resonance experiments. Asparagine C_γ assignments were then linked to side chain $H_{E/Z}$ proton pairs with an NH_2-filtered ^1H-^{13}C HSQC. Stereospecific assignments of asparagine H_E and H_Z resonances were determined from *J*-coupling values measured in E.COSY experiments (see Results) and were confirmed using proton-proton NOEs. All side chain assignments were made using Sparky [62]. Wild-type (except for stereospecific assignments), Y131F/D146L (YD), and L69A side chain assignments were inferred by comparison with C123N, and T49L spectra.

Chemical shift perturbations (CSPs) were calculated as the weighted Euclidean distance between wild-type and variant chemical shifts ($\delta_{H/N}^{WT}$, $\delta_{H/N}^{Var}$) using the equation

$$(1) \qquad CSP_{HN} = \sqrt{\frac{(\delta_H^{WT} - \delta_H^{Var})^2 + (0.14(\delta_N^{WT} - \delta_N^{Var}))^2}{2}}$$

where 0.14 is a weighting factor to normalize δ_N to δ_H [40].

^{15}N transverse relaxation rates with attenuated exchange contributions (R_2^*) were determined using 2D constant-time (CT) CPMG experiments [34]. For two-point relaxation dispersion experiments ν_{cpmg} was set to 200 Hz ($\nu_{cpmg} = 1/4\ \tau_{cp}$, $\tau_{cp} = 1.25$ ms) or 1 kHz ($\tau_{cp} = 250$ μs) to observe relaxation with maximal and minimal effects of chemical exchange, respectively. R_2^* values were determined using ν_{cpmg} of 1 kHz to measure

exchange-minimized relaxation of backbone and side-chain amides on the ps-ns timescale, with R_2^* values calculated using eq 1 from [34].

^{1}H-^{15}N NOE experiments were performed using the pulse sequence in [63]. The ratio between spectra with and without ^{1}H saturation was used to determine ^{1}H-^{15}N NOE effects.

2.4.4 Temperature coefficients.

Chemical shift values used to determine temperature coefficients for wild-type and all peripheral variants were obtained from ^{1}H, ^{15}N-HSQC spectra using a Bruker Avance 600 MHz spectrometer equipped with a cryoprobe. Additional chemical shift data for the T49L variant was obtained from TROSY spectra of an ^{15}N-labelled, 50% ^{2}H T49L sample using a Varian 800 MHz spectrometer equipped with a room temperature triple-resonance probe. HSQC spectra were collected at 10, 15, 20, and 30 °C; TROSY spectra were collected at 5 °C intervals from 10 to 35 °C. Peak assignments for each variant were determined by referencing NH_2 peaks from the 20 °C spectrum to the NH_2 assignments of wild-type, C123N, and T49L pp32; peaks could then be tracked by overlaying all spectra from the temperature series. Chemical shifts for resolvable amides were fit to a linear model. Amides that were unresolved at any temperature were excluded from the analysis.

2.4.5 Hydrogen exchange of asparagine side-chain NH_2 groups.

Hydrogen exchange rates from side-chain amides were measured using two types of experiments. To obtain an overall exchange rates for wild-type and variant pp32

constructs, we used NH_2-filtered HSQC experiments. This approach has the advantage that it suppresses slow-exchanging resonances from amide NH groups. In this experiment, decrease in NH_2 signal intensity results from exchange of either the H_Z or the H_E with a deuteron. As such, the apparent rate constant is the sum of the exchange rate constants for the H_Z and H_E protons. In addition, we used unfiltered ^{1}H-^{15}N HSQC spectra from the backbone hydrogen exchange studies of Dao et al [35] to resolve the exchange rates of the H_E and H_Z protons for wild-type pp32.

For the NH_2-filtered exchange measurements, exchange was initiated as described in [21]. A spin column was packed with 2 mL of pre-swollen G-25 fine Sephadex (GE Healthcare) and was washed three times with 2 mL water followed by three 2 mL washes with hydrogen exchange buffer (150 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, 100% D_2O, pH 6.8 (after accounting for D_2O)). Samples were applied to the equilibrated column, collected by centrifugation, and were immediately placed into a Bruker Avance 600 MHz spectrometer at 30 °C. NH_2-filtered HSQC spectra were recorded every 30 minutes for up to 4 hours. After 4 hours, spectra were recorded at longer intervals (hours or days); samples were immersed in a 30 °C water bath in between spectra.

Exchange rates for NHD species were obtained from data collected in [21]. Assignments were made by comparison to the wild-type NH_2-filtered HSQC spectrum determined in the present study. Since there is substantial overlap between the NH_2 and NHD peaks for both H_Z and H_E protons, peak heights were obtained by fitting cross peaks using a single two-dimensional lorentzian functions. The quality of fits was determined by

visual inspection of 1D slices in the ^{1}H and ^{15}N dimensions; when fitting with a single lorentzian results in systematic residuals, a second lorentizan was included in the fit.

The decay of NH_2-filtered NH_2 peak heights during hydrogen exchange experiments was fitted with a single-exponential decay using the equation

$$(2) \qquad I(t) = Ae^{-k_{ex}t} + B$$

where *I(t)* is the peak height at time *t*, *A* is the initial peak height (at *t = 0*) above the baseline *B* at *t* = 0, and k_{ex} is the rate constant for exchange. The build-up and subsequent decay of unfiltered asparagine 74 $(ND_E)H_Z$ peak heights was fitted with the double exponential eq 3 below. The build-up of unfiltered asparagine 98 $(ND_E)H_Z$ and $(ND_Z)H_E$ peak heights were globally fitted with eqs 3 and 4,

$$(3) \qquad I_Z(t) = A_Z(e^{-k_{ex,H_Z}t} - e^{-(k_{ex,H_E}+k_{ex,H_Z})t}) + B_Z$$

$$(4) \qquad I_E(t) = A_E(e^{-k_{ex,H_E}t} - e^{-(k_{ex,H_E}+k_{ex,H_Z})t}) + B_E$$

where $I_Z(t)$ and $I_E(t)$ are the peak heights for the $(ND_E)H_Z$ and $(ND_Z)H_E$ isotopomers respectively, k_{ex,H_E} and k_{ex,H_Z} are the rate constants for hydrogen exchange of H_E and H_Z respectively and are shared globally during the fit, and A_Z and A_E are the initial peak heights of the $(NH_E)H_Z$ and $(NH_Z)H_E$ isotopomers above the baselines B_Z and B_E, respectively. Protection factors (PFs) were calculated using the formula

$$(5) \qquad PF = \frac{k_i}{k_{ex}}$$

where k_i is the intrinsic exchange rate constant for an asparagine side chain and k_{ex} is a fitted rate constant from eqs 2, 3, or 3 and 4. The k_i value was determined from a linear

fit of the temperature dependence of rate constants for asparagine side chain exchange from previous studies. [44], [64], [65]. To account for potential sample degradation and spectrometer drift over the long exchange times of the wild-type and stabilizing variants (T49L and C123N), peak heights were normalized against non-exchangeable methyl peaks over the course of the experiment. For rapidly exchanging variants (L69A and YD, < 24 hours), normalization was not necessary.

For hydrogen exchange measurements on asparagine 74 in the L69A variant, we were only able to measure a few spectra before the NH_2 signal intensity decayed to the baseline. Thus, we fixed the *A* parameter from eq 2 due to a lack of data in the decay portion of the curve. *A* was fixed at the average value from the other asparagine 74 hydrogen exchange experiments.

2.5 Supplemental Figures

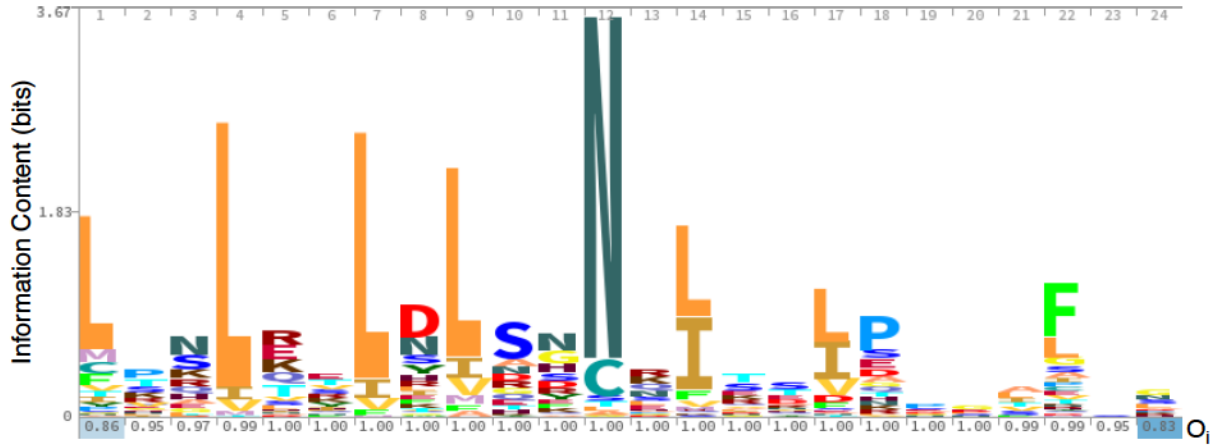


Figure S2.1. Sequence conservation in the LRR protein family. (A) HMM logo [66] for LRR_TYP and LRR families from the SMART database [67]. A starting set of 428,451 LRR sequences from the UniProt, Ensembl, or STRING databases was culled to retain sequences from the typical LRR subfamily (to which pp32 LRRs best align) by selecting for repeats from animals or fungi that were between 20 and 27 residues in length [19]. Using CD-Hit [68], sequences with greater than 80% identity were removed and the remaining sequences were aligned using MAFFT [69] with the gap opening and elongation penalties maximized. Occupancies (the fraction of sequences with gaps at position *i*; O_i) are tabulated beneath each letter stack. Positions with occupancies less than 50% are not shown. The decrease in occupancy for N- and C-terminal positions likely arises from variation in LRR length within the final data set.

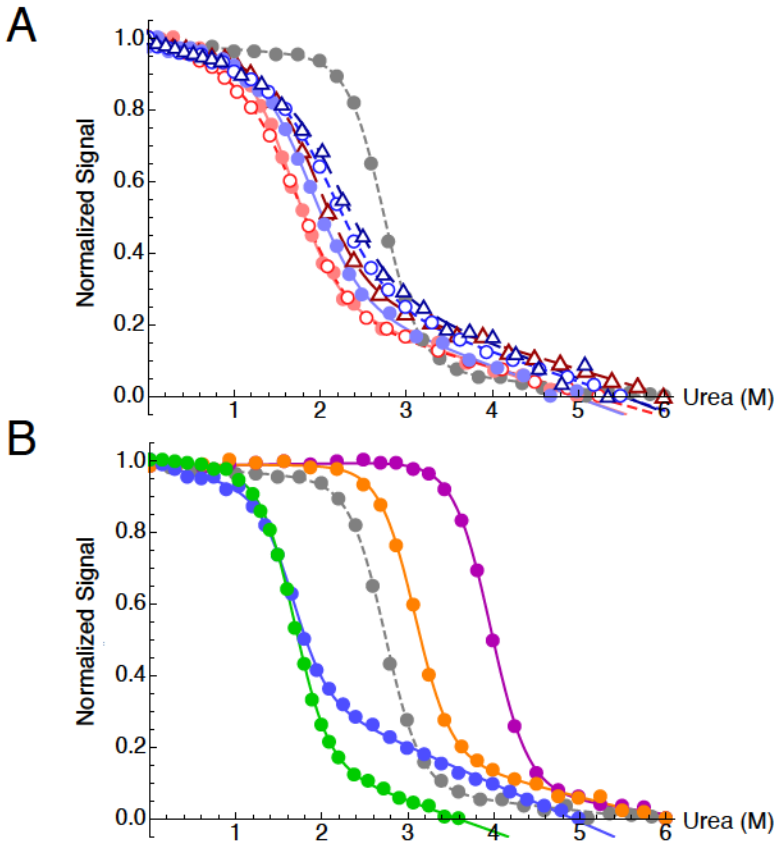


Figure S2.2. Urea-induced unfolding of Asn ladder and peripheral variants of pp32. Urea melts of asparagine ladder variants depicted as in Figure 1B, D. Transitions were monitored by CD at 220 nm and were fitted using a two-state model (curves). Data are reported in units of normalized ellipticity for asparagine ladder substitutions (A) and peripheral variants (B) by setting the highest and lowest CD values to one and zero respectively, and scaling all other points to these two limits. Conditions: ~3-5 μM protein, 20 mM $NaPO_4$, 150 mM NaCl, 0.1 mM TCEP, pH 7.8, 20 °C.

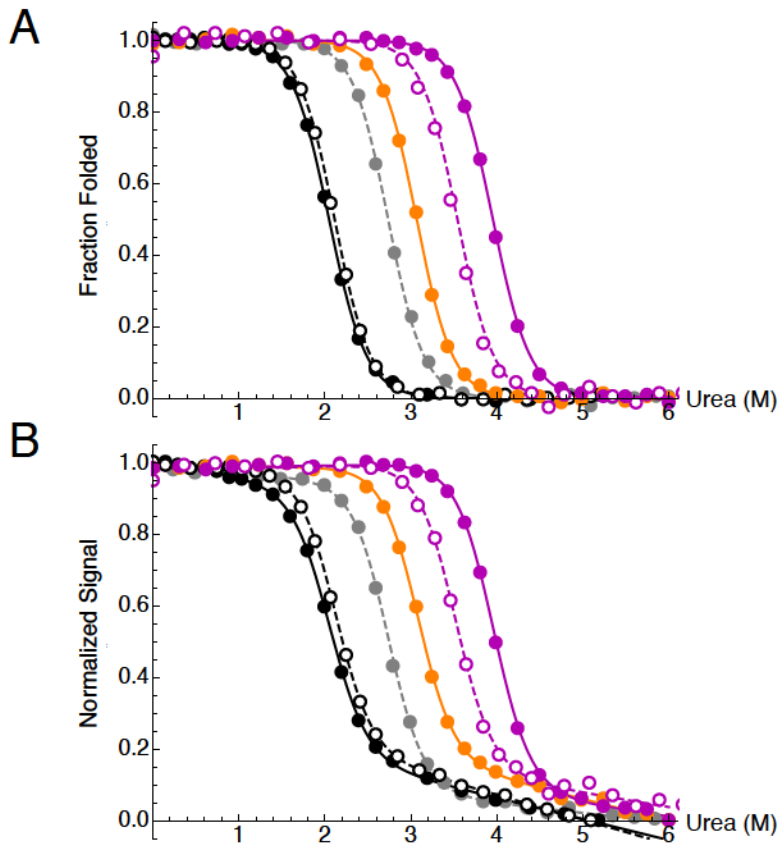


Figure S2.3. Urea-induced unfolding of Asn ladder extending and T49 substitutions in pp32. Urea denaturation of wild-type (dashed gray), N-terminal asparagine ladder extending substitutions (S27N, solid black; V52N, dashed black), C-terminal asparagine ladder extending substitutions (C123N, orange), and T49 substitutions (T49L, purple; T49V, dashed purple). Transitions were monitored by CD at 220 nm and were fitted using a two-state model (curves). Data and curves are transformed to (A) fraction folded using fitted baseline parameters or (B) normalized ellipticity by setting the highest and lowest CD values to one and zero respectively, and scaling all other points to these two limits (B). Conditions: ~3-5 μM protein, 20 mM $NaPO_4$, 150 mM NaCl, 0.1 mM TCEP, pH 7.8, 20 °C.

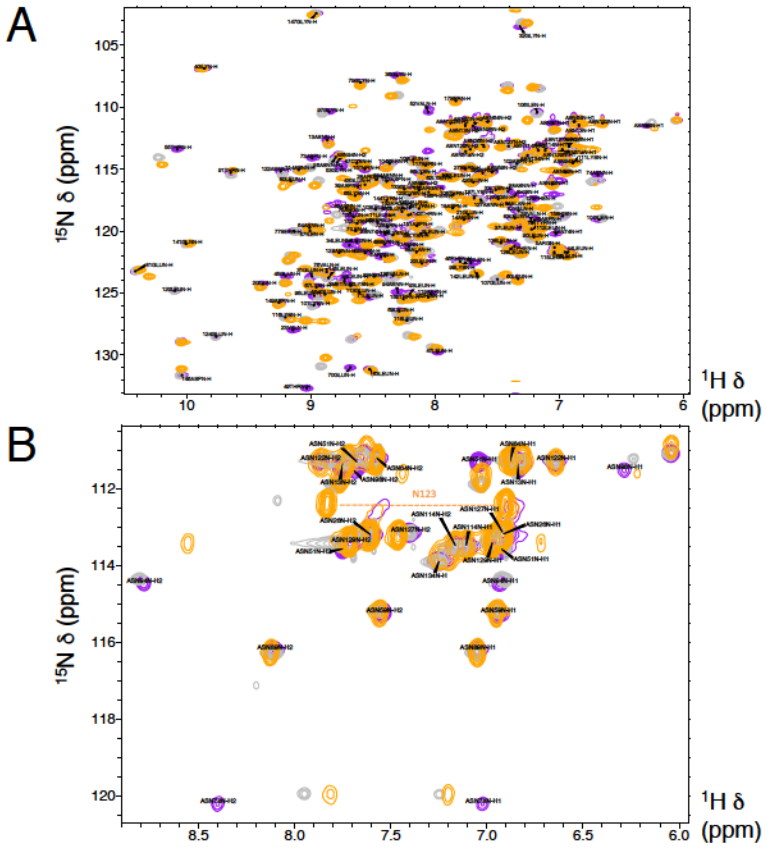


Figure S2.4. Backbone NH and NH_2-filtered HSQC spectra of pp32 variants. (A) Overlay of ^{1}H-^{15}N HSQC spectra for select pp32 variants (wild-type, gray; C123N, orange; T49L, purple). Assignments are shown for the T49L variant, as this protein had the largest number of assignable resonances. (B) Overlay of NH_2-filtered HSQC spectra colored and labeled as in (A). The new N123 peak in the C123N variant is indicated with an orange dashed line and label. Conditions: ~600-800 μM protein, 10% D_2O, 20 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, pH 6.8, 20 °C.

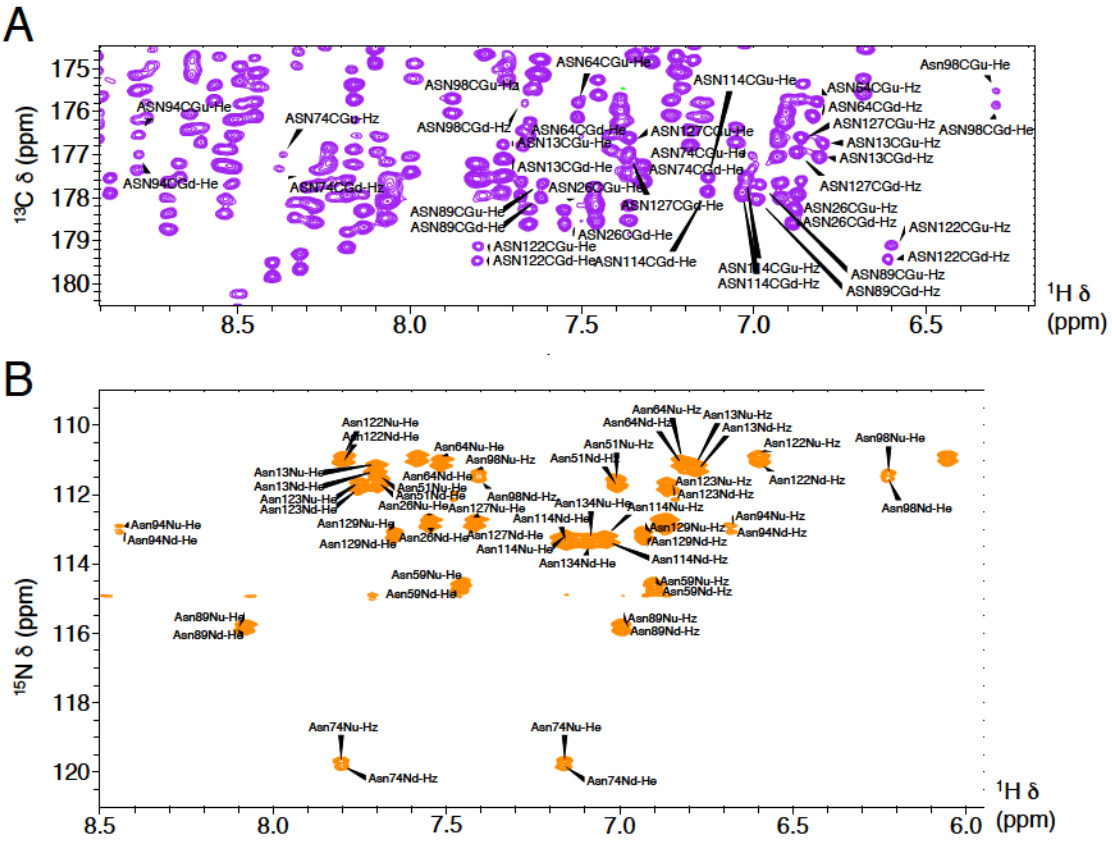


Figure S2.5. Full HNCO E.COSY spectra from T49L and C123N variants. (A) The NH_2 region of the HNCO E.COSY spectrum of the T49L variant of pp32. In this experiment, C_β spin states are not perturbed; therefore, cross peaks in ω_1 and ω_2 couple to the same C_β spin state (*i.e.*, C_γ C_β^α and $H_{E/Z}$ C_β^α). The large positive $^1J(C_\beta\text{-}C_\gamma)$ coupling allows the small $^3J(H_Z\text{-}C_\beta)$ (5-10 Hz) and $^3J(H_E\text{-}C_\beta)$ (effectively 0 Hz) to be determined, thereby assigning the asparagine H_Z and H_E protons. Resolved asparagine side chain peaks are labeled; unlabeled peaks are residual signals from backbone amides. (B) Full $C_\gamma O$-coupled NH_2-HSQC E.COSY for the C123N variant of pp32. As in (A), the C_γ spin states are unperturbed. The large $^1J(N_{\delta 2}\text{-}C_\gamma)$ allows the small positive $^2J(H_E\text{-}C_\gamma)$ (~1-5 Hz) and negative $^2J(H_Z\text{-}C_\gamma)$ (~ -1 to -5 Hz) to be determined, thereby assigning the asparagine H_Z and H_E protons. Nu = upfield ^{15}N doublet peak, Nd = downfield ^{15}N doublet peak. Conditions: ~600 μM protein, 50% D_2O, 20 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, pH 6.8 (after adjustment for D_2O), 35 °C.

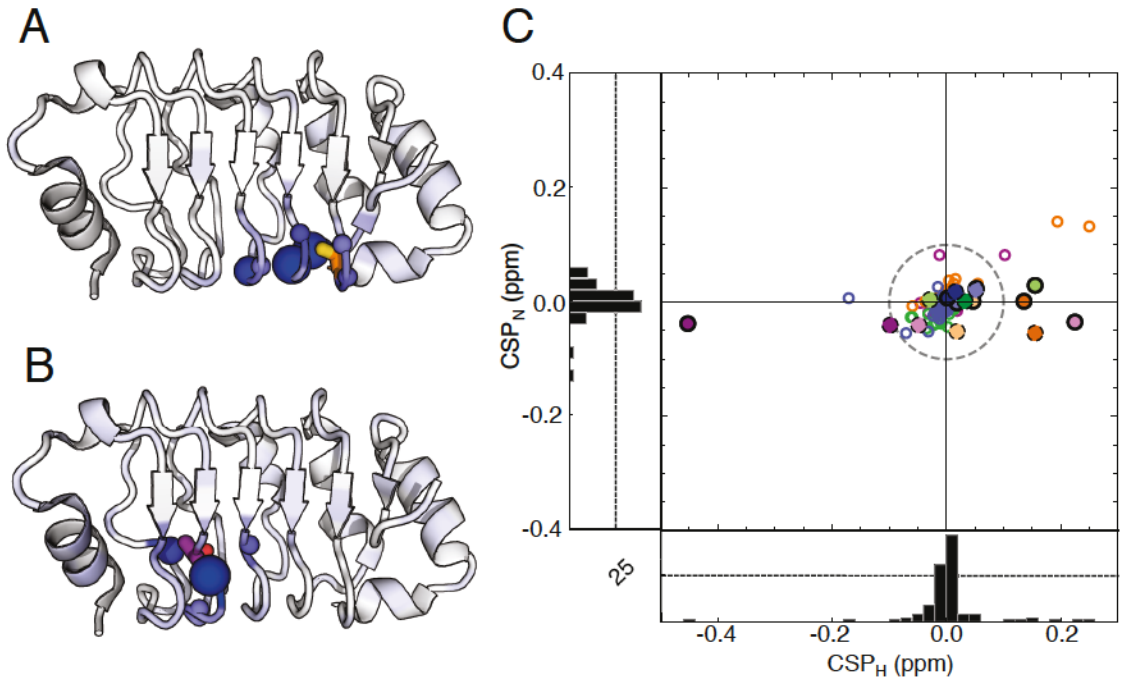


Figure S2.6. Backbone amide and asparagine side chain chemical shift perturbations in pp32 variants. (A) C123N CSPs mapped to the pp32 structure (PDB ID: 4XOS). The magnitude of the CSP at each residue is indicated by the blue shading intensity. Large spheres show residues where CSPs exceed $3\sigma_{CSP}$ (where σ_{CSP} is the standard deviation in chemical shift perturbation values), medium spheres show residues where $2\sigma_{CSP} < CSP < 3\sigma_{CSP}$, and small spheres show residues where $1\sigma_{CSP} < CSP < 2\sigma_{CSP}$. (B) T49L CSPs mapped to the pp32 structure. Color mapping and sphere size cutoffs are as in (A). (C) ^{1}H-^{15}N CSPs for all resolvable side chain protons. Non-ladder asparagines are shown as empty circles. Ladder asparagines 74 and 98 are shown as filled circles with solid and dashed outlines, respectively; H_Z and H_E have dark and light shading, respectively. T49L, purple; L69A, green; C123N, orange; YD, blue. As in an HSQC spectrum, quadrants to the left indicate increased ^{1}H deshielding in the variants compared to wild-type pp32, and lower quadrants indicate increased ^{15}N deshielding. The dashed gray circle is one σ from the median. All distances are measured in ^{1}H units of ppm with a scaling factor of 0.14 in the ^{15}N-dimension. Histograms represent the number of side chain protons with a particular CSP_N or CSP_H value; the dashed line across each histogram shows a count of 25 CSPs. Conditions: ~600 μM protein, 20 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, pH 6.8, 20 °C.

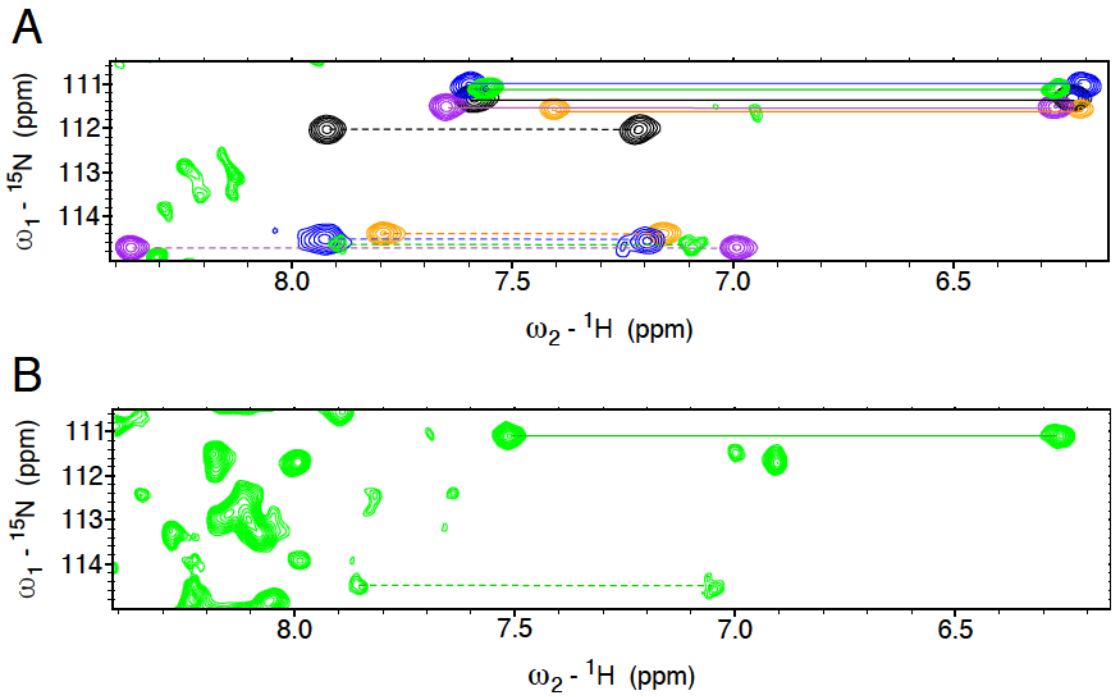


Figure S2.7. NH_2-filtered HSQC spectra for partly exchanged samples of wild-type pp32 and variants. (A) NH_2-filtered HSQC spectra from the first hydrogen exchange time point for wild-type pp32 and variants. Peaks are folded in the ^{15}N dimension as a result of a decreased spectral width to shorten data acquisition time. Colors are as in Figure 2; dashed lines connect N74 resonances, solid lines connect N98 resonances. L69A peaks are from a hydrogen exchange experiment at 20°C with reduced background noise (see B below for 30°C experiment). (B) NH_2-filtered HSQC spectra from hydrogen exchange first time point for L69A at 30°C. Assignments of the N74 resonances of L69A were corroborated by comparison to an NH_2-filtered ^1H-^{15}N HSQC spectrum with a higher signal-to-noise ratio. Conditions: ~800 μM protein, 100% D_2O, 150 mM $NaPO_4$, 50 mM NaCl, 0.1 mM TCEP, 30 °C (20 °C for L69A in (A)) pH 6.8 (after accounting for D_2O).

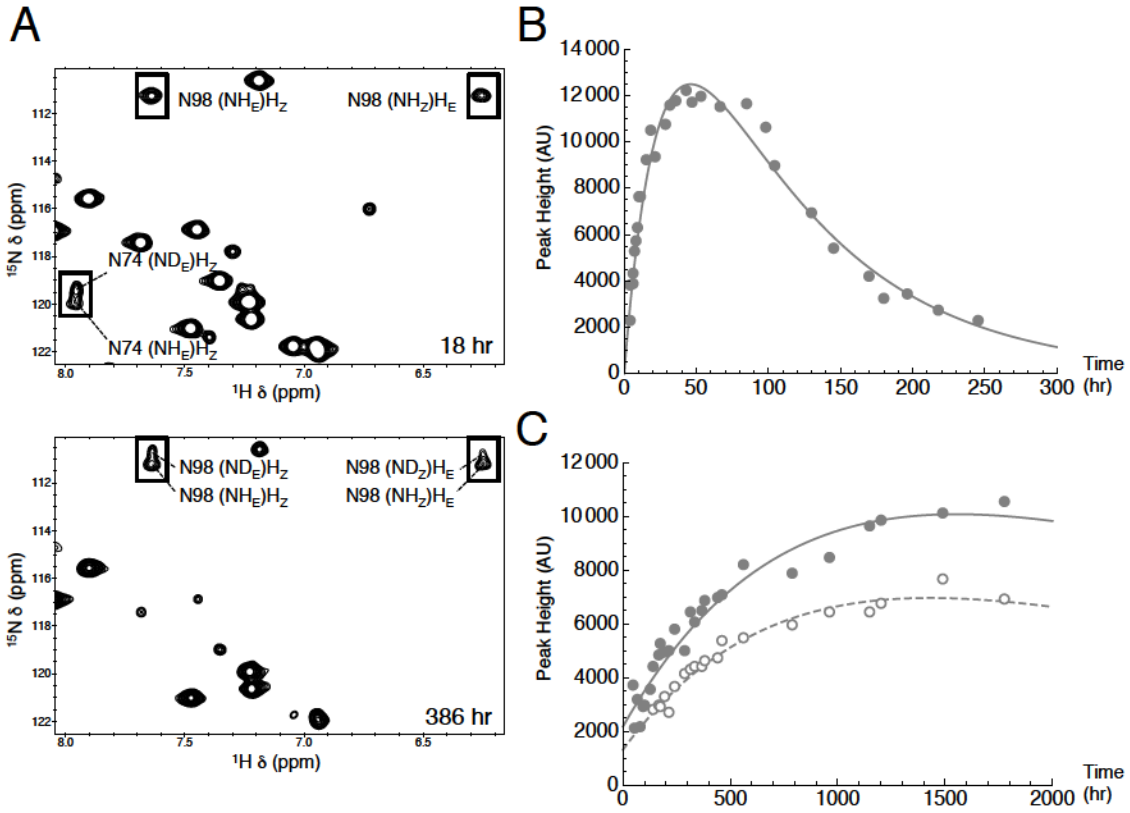


Figure S2.8. Hydrogen exchange of ladder asparagine side chains from unfiltered ^{1}H-^{15}N HSQC spectra. (A) Unfiltered ^{1}H-^{15}N HSQC spectra of wild-type pp32 after 18 (top) and 386 hours (bottom). NH_2 and NHD peaks are resolved by a slight shift of the NHD species to lower frequency in the ^{15}N dimension. (B) N74 $(ND_E)H_Z$ peak intensity as a function of exchange time for wild-type pp32, fitted with eq 3 (solid curve). (C) N98 $(ND_E)H_Z$ (filled circles) and $(ND_Z)H_E$ (empty circles) peak height as a function of exchange time for wild-type pp32 globally fitted with eqs 3 and 4 (solid and dashed curves). Unlike N74 (B), the individual exchange profiles for the N98 NHD species only show the build-up phase due to the slow exchange at N98. However, the $(N_DZ)H_E$ and $(ND_E)H_Z$ build-ups report on different rate constants ($k_{ex,HE}$ and $k_{ex,HZ}$, respectively), allowing both constants to be determined accurately from the 2000 hour exchange profile. Conditions: 800 μM protein, 100% D_2O, 20 mM $NaPO_4$, 50 mM NaCl, 0.2 mM TCEP, 20 °C, pH 6.7 (after accounting for D_2O).

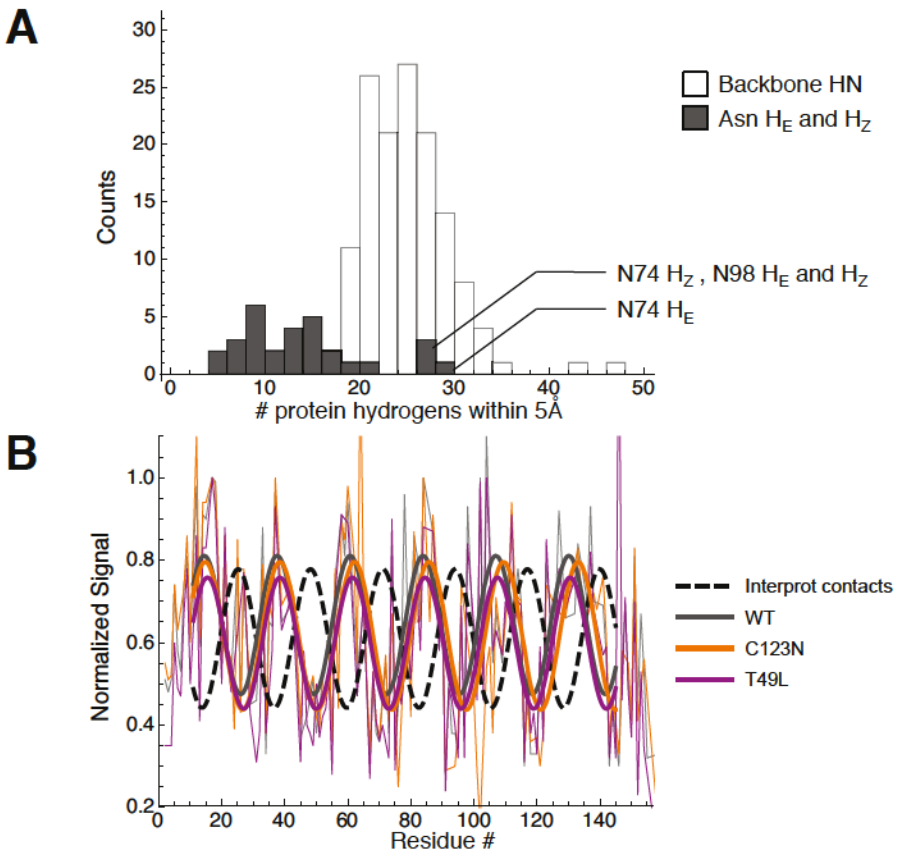


Figure S2.9. Correlation between interproton contacts and ^{1}H-^{15}N HSQC peak heights. (A) Histogram of the number of protein hydrogens (excluding waters) within 5 Å of backbone amide hydrogens (white bars) and asparagine side chain H_E and H_Z (gray bars) in the high-resolution crystal structure of pp32 (PDB ID: 4XOS). (B) Comparison of cosine function fits to HSQC peak heights (solid curves) and the number of protein hydrogens within 5 Å (dashed curve) in pp32. Peak heights (thin lines) are normalized to the height of the NH peak of residue 17 (one of the five most intense peaks in all variants). The dashed cosine function is fit to the normalized number of interproton contacts to backbone amide hydrogens from (A) as a function of position and is scaled and shifted along the y-axis to match the mean and amplitude of the pp32 variants. All fits were limited to residues within the five leucine-rich repeats (residues 11-145) using the function $I(x) = a*\cos(2\pi*(x-c/n))+d$, where I is the intensity, a is the amplitude, x is the residue number, c is a phase term, n is the number of residues per repeating unit, and d is a vertical offset.

2.6 References

[1] C. L. Worth and T. L. Blundell, “On the evolutionary conservation of hydrogen bonds made by buried polar amino acids: The hidden joists, braces and trusses of protein architecture,” *BMC Evol. Biol.*, 2010.

[2] X. Yang, S. V. Kathuria, R. Vadrevu, and C. R. Matthews, “βα-Hairpin clamps brace βαβ modules and can make substantive contributions to the stability of TIM barrel proteins,” *PLoS One*, 2009.

[3] B. Basanta *et al.*, “Introduction of a polar core into the de novo designed protein Top7,” *Protein Sci.*, vol. 00, pp. 1299–1307, 2016.

[4] M. F. Perutz, “Electrostatic effects in proteins,” *Science (80-.).*, vol. 201, pp. 1187–1191, 1978.

[5] B. C. Cunningham and J. A. Wells, “High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis,” *Science (80-.).*, 1989.

[6] L. C. Wheeler, S. A. Lim, S. Marqusee, and M. J. Harms, “The thermostability and specificity of ancient proteins,” *Current Opinion in Structural Biology*. 2016.

[7] Z. R. Sailer and M. J. Harms, “Molecular ensembles make evolution unpredictable,” *Proc. Natl. Acad. Sci.*, 2017.

[8] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl, “Using information theory to search for co-evolving residues in proteins,” *Bioinformatics*, 2005.

[9] I. Anishchenko, S. Ovchinnikov, H. Kamisetty, and D. Baker, “Origins of coevolution between residues distant in protein 3D structures,” *Proc. Natl. Acad. Sci.*, 2017.

[10] S. Nakano *et al.*, “Benchmark analysis of native and artificial NAD+-dependent enzymes generated by a sequence-based design method with or without phylogenetic data,” *Biochemistry*, vol. 57, pp. 3722–3732, 2018.

[11] V. D. Goyal and T. J. Magliery, “Phylogenetic spread of sequence data affects fitness of SOD1 consensus enzymes: Insights from sequence statistics and structural analyses,” *Proteins Struct. Funct. Bioinforma.*, 2018.

[12] F. Morcos *et al.*, “Direct-coupling analysis of residue coevolution captures native contacts across many protein families,” *Proc. Natl. Acad. Sci.*, 2011.

[13] M. J. Harms and J. W. Thornton, “Historical contingency and its biophysical basis in glucocorticoid receptor evolution,” *Nature*, 2014.

[14] V. H. Salinas and R. Ranganathan, “Coevolution-based inference of amino acid interactions underlying protein function,” *Elife*, 2018.

[15] G. A. Faiman and A. Horovitz, “On the choice of reference mutant states in the application of the double-mutant cycle method,” *Protein Engineering, Design and Selection*. 1996.

[16] W. A. Baase, L. Liu, D. E. Tronrud, and B. W. Matthews, “Lessons from the lysozyme of phage T4,” *Protein Science*. 2010.

[17] A. S. Canale, P. A. Cote-Hammarlof, J. M. Flynn, and D. N. Bolon, “Evolutionary mechanisms studied through protein fitness landscapes,” *Current Opinion in Structural Biology*. 2018.

[18] D. S. Eisenberg and M. R. Sawaya, “Structural Studies of Amyloid Proteins at the

Molecular Level," *Annu. Rev. Biochem.*, 2017.
[19] A. V. Kajava, "Structural diversity of leucine-rich repeat proteins," *J. Mol. Biol.*, vol. 277, no. 3, pp. 519–527, 1998.
[20] N. Courtemanche and D. Barrick, "The Leucine-Rich Repeat Domain of Internalin B Folds along a Polarized N-Terminal Pathway," *Structure*, vol. 16, no. 5, pp. 705–714, 2008.
[21] T. P. Dao, A. Majumdar, and D. Barrick, "Highly polarized C-terminal transition state of the leucine-rich repeat domain of PP32 is governed by local stability," *Proc. Natl. Acad. Sci.*, 2015.
[22] A. G. Evdokimov, D. E. Anderson, K. M. Routzahn, and D. S. Waugh, "Unusual molecular architecture of the Yersinia pestis cytotoxin YopM: A leucine-rich repeat protein with the shortest repeating unit," *J. Mol. Biol.*, 2001.
[23] B. Kobe and J. Deisenhofer, "Proteins with leucine-rich repeats," *Curr. Opin. Struct. Biol.*, 1995.
[24] T. Huyton and C. Wolberger, "The crystal structure of the tumor suppressor protein pp32 (Anp32a): Structural insights into Anp32 family of proteins," *Protein Sci.*, 2007.
[25] S. Zamora-Caballero, L. Šiaučiunaite-Gaubard, and J. Bravo, "High-resolution crystal structure of the leucine-rich repeat domain of the human tumour suppressor PP32A (ANP32A)," *Acta Crystallogr. Sect. F Struct. Biol. Commun.*, 2015.
[26] T. P. Dao, A. Majumdar, and D. Barrick, "Capping motifs stabilize the leucine-Rich repeat protein PP32 and rigidify adjacent repeats," *Protein Sci.*, 2014.
[27] M. J. Fossat *et al.*, "High-Resolution Mapping of a Repeat Protein Folding Free Energy Landscape," *Biophys. J.*, 2016.
[28] A. Matilla and M. Radrizzani, "The Anp32 family of proteins containing leucine-rich repeats," *Cerebellum*. 2005.
[29] K. A. Jenkins *et al.*, "The consequences of cavity creation on the folding landscape of a repeat protein depend upon context," *Proc. Natl. Acad. Sci.*, 2018.
[30] T. Harsch, P. Schneider, B. Kieninger, H. Donaubauer, and H. R. Kalbitzer, "Stereospecific assignment of the asparagine and glutamine sidechain amide protons in proteins from chemical shift analysis," *J. Biomol. NMR*, 2017.
[31] F. Löhr and H. Rüterjans, "H2NCO-E. COSY, a Simple Method for the Stereospecific Assignment of Side-Chain Amide Protons in Proteins," *J. Magn. Reson.*, 1997.
[32] M. Cai, Y. Huang, and G. M. Clore, "Accurate orientation of the functional groups of asparagine and glutamine side chains using one-and two-bond dipolar couplings [25]," *Journal of the American Chemical Society*. 2001.
[33] F. Cordier and S. Grzesiek, "Direct observation of hydrogen bonds in proteins by interresidue (3h)J(NC') scalar couplings [3]," *Journal of the American Chemical Society*. 1999.
[34] F. A. A. Mulder, N. R. Skrynnikov, B. Hon, F. W. Dahlquist, and L. E. Kay, "Measurement of slow (μs-ms) time scale dynamics in protein side chains by 15N relaxation dispersion NMR spectroscopy: Application to Asn and Gln residues in a

cavity mutant of T4 lysozyme," *J. Am. Chem. Soc.*, 2001.
[35] Q. Gong and R. Ishima, "15N-{1H} NOE experiment at high magnetic field strengths," *J. Biomol. NMR*, 2007.
[36] N. J. Baxter and M. P. Williamson, "Temperature dependence of 1H chemical shifts in proteins," *J. Biomol. NMR*, 1997.
[37] F. Cordier and S. Grzesiek, "Temperature-dependence of protein hydrogen bond properties as studied by high-resolution NMR," *J. Mol. Biol.*, 2002.
[38] T. Cierpicki, I. Zhukov, R. A. Byrd, and J. Otlewski, "Hydrogen bonds in human ubiquitin reflected in temperature coefficients of amide protons," *J. Magn. Reson.*, 2002.
[39] J. Hong, Q. Jing, and L. Yao, "The protein amide1HNchemical shift temperature coefficient reflects thermal expansion of the N-H···O=C hydrogen bond," *J. Biomol. NMR*, 2013.
[40] J. H. Tomlinson and M. P. Williamson, "Amide temperature coefficients in the protein G B1 domain," *J. Biomol. NMR*, 2012.
[41] T. Cierpicki and J. Otlewski, "Amide proton temperature coefficients as hydrogen bond indicators in proteins," *J. Biomol. NMR*, 2001.
[42] J. J. Skinner, W. K. Lim, S. Bédard, B. E. Black, and S. W. Englander, "Protein dynamics viewed by hydrogen exchange," *Protein Sci.*, 2012.
[43] P. Rajagopal, B. E. Jones, and R. E. Klevit, "Solvent exchange rates of side-chain amide protons in proteins," *J. Biomol. NMR*, 1998.
[44] N. R. Krishna *et al.*, "Primary Amide Hydrogen Exchange in Model Amino Acids: Asparagine, Glutamine, and Glycine Amides," *J. Am. Chem. Soc.*, 1982.
[45] H. M. Kim *et al.*, "Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran," *Cell*, vol. 130, no. 5, pp. 906–917, 2007.
[46] E. Tuchsen and C. Woodward, "Hydrogen Exchange of Primary Amide Protons in Basic Pancreatic Trypsin Inhibitor: Evidence for NH2 Group Rotation in Buried Asparagine Side Chains," *Biochemistry*, 1987.
[47] G. I. Makhatadze, K. -S Kim, C. Woodward, and P. L. Privalov, "Thermodynamics of bpti folding," *Protein Sci.*, 1993.
[48] A. Liu, Z. Lu, J. Wang, L. Yao, Y. Li, and H. Yan, "NMR detection of bifurcated hydrogen bonds in large proteins," *J. Am. Chem. Soc.*, 2008.
[49] M. R. Preimesberger *et al.*, "Direct NMR detection of bifurcated hydrogen bonding in the α-helix N-caps of ankyrin repeat proteins," *J. Am. Chem. Soc.*, 2015.
[50] E. Kloss and D. Barrick, "Thermodynamics, Kinetics, and Salt dependence of Folding of YopM, a Large Leucine-rich Repeat Protein," *J. Mol. Biol.*, vol. 383, no. 5, pp. 1195–1209, 2008.
[51] E. Kloss and D. Barrick, "C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein," *Protein Sci.*, vol. 18, no. 9, pp. 1948–1960, 2009.
[52] V. J. LiCata and G. K. Ackers, "Long-Range, Small Magnitude Nonadditivity of Mutational Effects in Proteins," *Biochemistry*, 1995.
[53] E. Di Cera, *Thermodynamic theory of site-specific binding processes in biological macromolecules*. Cambridge University Press, 1995.

[54] D. D. Krantz, R. Zidovetzki, B. L. Kagan, and S. L. Zipursky, “Amphipathic β structure of a leucine-rich repeat peptide,” *J. Biol. Chem.*, 1991.
[55] N. J. Gay, L. C. Packman, M. A. Weldon, and J. C. J. Barna, “A leucine-rich repeat peptide derived from the Drosophila Toll receptor forms extended filaments with a β-sheet structure,” *FEBS Lett.*, 1991.
[56] M. F. Jeng and S. W. Englander, “Stable submolecular folding units in a non-compact form of cytochrome c,” *J. Mol. Biol.*, 1991.
[57] C. N. Pace, “Determination and Analysis of Urea and Guanidine Hydrochloride Denaturation Curves,” *Methods Enzymol.*, 1986.
[58] T. O. Street, N. Courtemanche, and D. Barrick, “Protein Folding and Stability Using Denaturants,” *Methods Cell Biol.*, vol. 84, no. 07, pp. 295–325, 2008.
[59] J. E. Masse and R. Keller, “AutoLink: Automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic,” *J. Magn. Reson.*, 2005.
[60] M. S. Chimenti, C. A. Castañeda, A. Majumdar, and B. García-Moreno E., “Structural Origins of High Apparent Dielectric Constants Experienced by Ionizable Groups in the Hydrophobic Core of a Protein,” *J. Mol. Biol.*, 2010.
[61] L. P. McIntosh, E. Brun, and L. E. Kay, “Stereospecific assignment of the NH2 resonances from the primary amides of asparagine and glutamine side chains in isotopically labeled proteins,” *J. Biomol. NMR*, 1997.
[62] T. D. Goddard and D. G. Kneller, “SPARKY 3.” University of California San Francisco.
[63] L. E. Kay, D. A. Torchia, and A. Bax, “Backbone Dynamics of Proteins As Studied by 15N Inverse Detected Heteronuclear NMR Spectroscopy: Application to Staphylococcal Nuclease,” *Biochemistry*, 1989.
[64] R. S. Molday, S. W. Englander, and R. G. Kallen, “Primary Structure Effects on Peptide Hydrogen Exchange,” *Biochemistry*, 1972.
[65] Y. Bai, J. S. Milne, L. Mayne, and S. W. Englander, “Primary structure effects on peptide group hydrogen exchange,” *Proteins Struct. Funct. Bioinforma.*, 1993.
[66] T. J. Wheeler, J. Clements, and R. D. Finn, “Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models,” *BMC Bioinformatics*, 2014.
[67] I. Letunic and P. Bork, “20 years of the SMART protein domain annotation resource,” *Nucleic Acids Res.*, 2018.
[68] W. Li and A. Godzik, “Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, 2006.
[69] K. Katoh, J. Rozewicki, and K. D. Yamada, “MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization,” *Brief. Bioinform.*, 2017.

CHAPTER 3 – Single repeat resolution of a consensus LRR protein using a nearest-neighbor model.

3.1 Introduction

Studying how constellations of interacting residues contribute to the phenomenon of cooperative folding in proteins remains challenging. In globular proteins, residues form contacts across a wide spectrum of primary sequence distances. These long-range contacts make it difficult to break globular proteins into their individual structural components [1]. Powerful techniques such as hydrogen exchange monitored via NMR or mass spectrometry [2] address these issues by resolving protection of individual amides. However, hydrogen exchange techniques are limited by proton exchange behavior and the natural folding path of the protein. If a proton does not exchange in an EX2 mechanism, the rate of exchange cannot be converted into a ΔG°. Likewise, if a domain does not fold first in a multi-domain protein, then its measured stability will always reflect its interaction with its environment as well as its intrinsic stability. These problems are not insurmountable but they can complicate studies trying to isolate the contributions of individual domains.

To break a protein into individual domains and measure the stability and coupling between the domains, the protein must be simplified. One way to achieve this is to reduce the complexity of the contacts. Repeat proteins have fewer long-range contacts than globular proteins and are composed of arrays of repeats with near-identical secondary structure and high sequence similarity [1]. The modular nature of repeat proteins makes it easier to isolate particular structural features. In addition to the simplicity of their contact

order, repeat proteins are also amenable to consensus design [3]. The ability to convert heterogeneous arrays of protein repeats into homogenous ones has resulted in detailed analyses of a number of consensus repeat proteins [4]–[9]. These analyses have explored a range of complex biological phenomena including folding pathways [10], polar networks [11], and fractured interfaces [7].

Of particular relevance to this thesis is a study of consensus LRR (cLRR) arrays [9]. This study is the only to date in which the folding thermodynamics of a β-sheet-containing repeat array has been characterized and analyzed using a nearest-neighbor model. Nearest-neighbor analysis of cLRR constructs revealed that the interfaces between cLRR repeats are the most favorable of any consensus protein studied in our lab [9]. The novelty of this study likely results from the difficulty in creating a soluble, foldable β-sheet repeat array. In [9], several different permutations of repeats and capping sequences were tried unsuccessfully before arriving at the design shown in Figure 3.1. The repeating unit in this design is composed of a pair of repeats referred to as A and B repeats. The A and B repeats differ at solvent-exposed polar or charged positions (Figure 3.1B) to avoid electrostatic repulsion between them [9]. In this analysis, the combined A and B repeats are referred to as a paired-repeat (R) and is termed the paired-repeat model to differentiate it from an analysis that can resolve folding parameters for the individual repeats (single-repeat model).

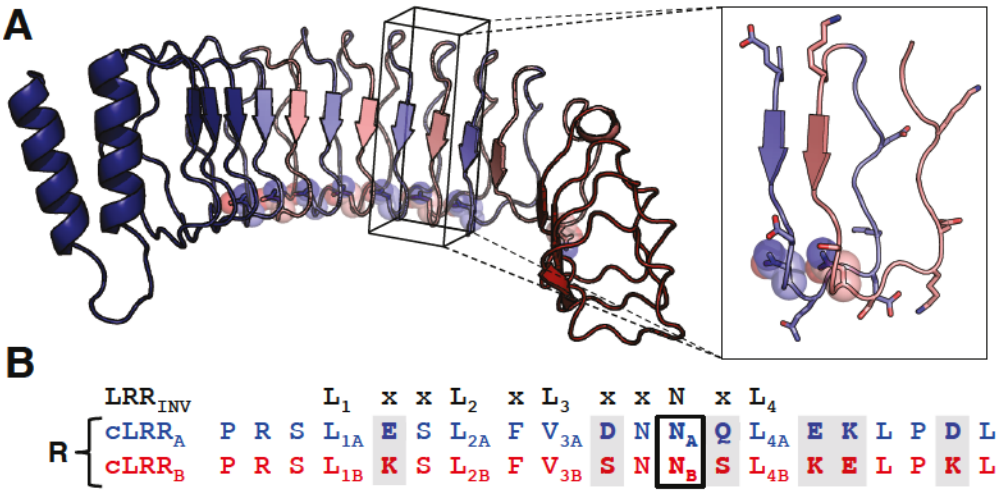


Figure 3.1. Consensus LRR structure and sequence. (A) Homology models of cLRR NR_4C structure from SWISS-MODEL webserver [12]. The model is composed of an N-cap (dark blue), four A-B repeat pairs (light blue and light red, respectively) repeats, and a C-cap (dark red). The putative asparagine ladder is shown in sticks and spheres. The inset shows a single A-B repeat pair with charge alternating sites shown in sticks and the asparagine ladder position shown in sticks and spheres. (B) Sequences of invariant LRR region (black), cLRR repeat A (blue), and B (red). Charge alternating positions and polar substitutions are highlighted in gray and the conserved asparagine position is highlighted with a black rectangle. As the cLRR sequence has a valine at the L_3 position, it is called V_3 when referring to the bacterial subfamily during the sequence analysis.

Although the paired-repeat approach determines the free energy of the interface between the B and A repeats, it cannot dissect the intrinsic stabilities of individual repeats or the A:B interfacial stability. To attempt to directly measure these intrinsic and interfacial stabilities and to quantify the energetic effects of A and B substitutions, I have generated additional constructs and conducted a single-repeat nearest-neighbor analysis to resolve parameters for the A and B cLRR repeats. The microscopic parameters can then be used for in-depth analysis of the effects from substitutions to key residues such as the asparagine ladder (see Chapter 4).

3.2 Results

3.2.1 Reconstructing the bacterial LRR subfamily.

The large number of LRR protein entries in sequence databases like Pfam [13] and SMART [14] provide a wealth of information for determining consensus sequences and pairwise sequence correlations. The SMART database is particularly useful for such analysis, since it breaks LRR proteins into individual LRRs. Here, I mined the SMART database (version 8.0) for all LRR sequences and filtered these sequences to identify entries matching the bacterial LRR subfamily (20 residue length, from the bacteria taxon).

The HMM logo of the MSA from the bacterial subfamily of LRRs aggregated here is similar to the dataset used to construct the cLRR sequence (Figure S3.1) [9]. Notable differences between the two profiles are a reduction in conservation of valine and higher frequency of cysteine at position nine, and a decrease in conservation at the asparagine ladder position in the MSA generated here. This apparent decrease in conservation may derive from removal of redundant sequences from this sequence set.

3.2.2 Pairwise couplings in the cLRR sequence.

Since previous studies have shown substantial inter-repeat coupling in LRR proteins (see Chapter 2) [15], individual LRR sequences were converted into pairs using sequence header information to match adjacent LRR repeats. This final set of sequence pairs was clustered to remove redundant sequences to create the LRR_BAC_PAIR sequence set (Figure S3.1A). To identify co-evolving residue pairs within or between

repeats, a mutual information (MI) matrix was constructed and modified using the average product correction procedure (Figure 3.2) [16]. The largest source of inter-repeat MI is between positions 9 and 29, which are identical positions in adjacent repeats ($V_{3A,B}$ in Figure 3.1). There is a slight preference for a valine at this position in the LRR_BAC_PAIR sequence set, but cysteine is also observed with high frequency. The high mutual information score between positions V_{3A} and V_{3B} results from the fact that CC and VV are the most common residue pairs, whereas VC and CV are very infrequent (Figure 3.3A).

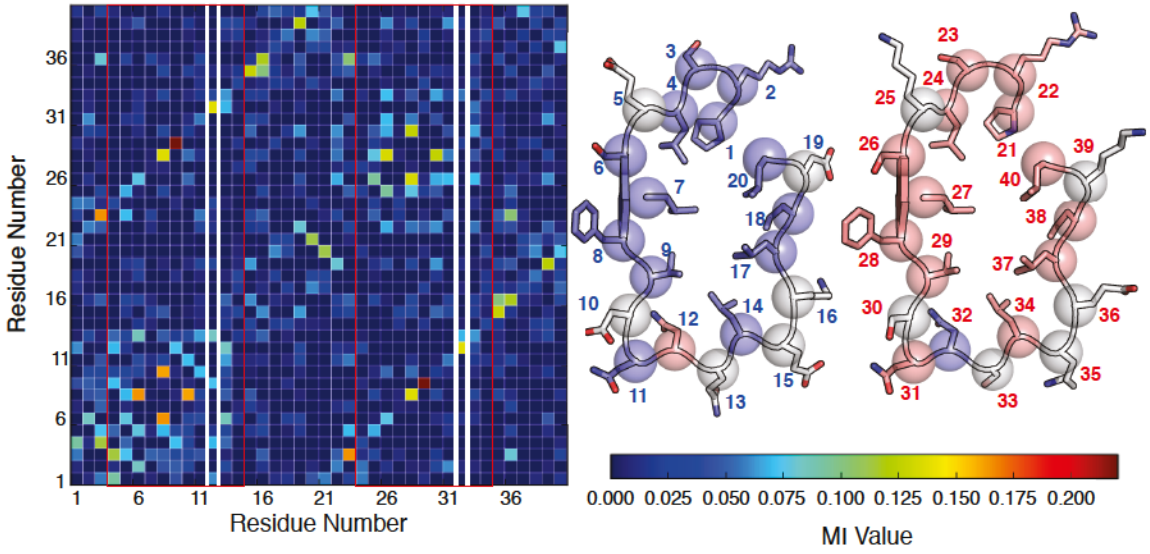


Figure 3.2. Mutual information for paired sequence from LRR bacterial subfamily. (B) Mutual information matrix for repeat pairs from the bacterial LRR subfamily. The red rectangles identify the conserved LRR pattern (LxxLxLxxNxL) with thick white rectangle indicating the conserved asparagine position. Mutual information values in the scale bar are from equation 3.1. Cartoon to the right shows the amino acid at each position for the pair of cLRR repeats. Repeats and numbering are colored as in Figure 3.1 with side chains shown with sticks and C_αs highlighted by spheres. Charge alternating positions are in gray and the conserved asparagine is in red (repeat A) or blue (repeat B).

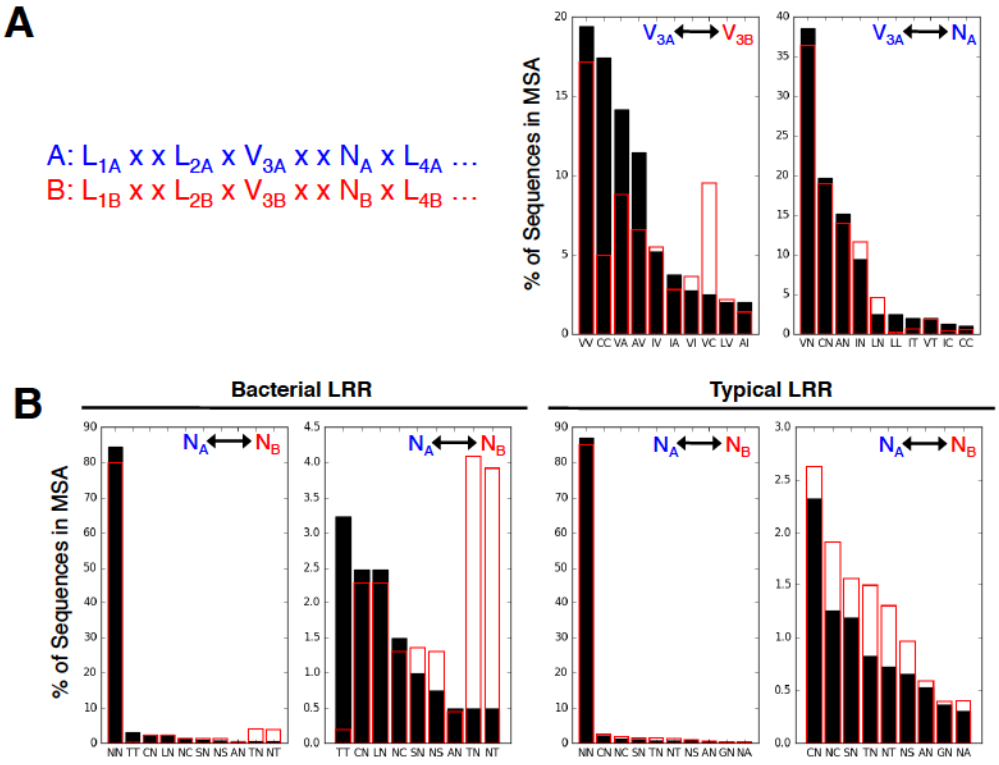


Figure 3.3. Frequencies of amino acid pairs within invariant LRR sequence. Bar charts showing frequencies of top ten residue pairs between two positions. In each plot, the two positions of the residue pair indicated by the text in the upper right hand corner of the plot and are labeled according to the scheme in (A). Black bars represent the joint probability observed for a residue pair ($p(n_1, n_2)$) while red outlines represent multiplication of the independent probability of each residue ($p(n_1) * p(n_2)$). Bar heights reflect the observed percentage of sequences within the MSA with a particular residue pair (black) or what would be expected from the independent probabilities (red). (A) Bacterial subfamily LRR invariant label scheme as in Figure 3.1B and bar charts for frequencies of residue pairs between the V_3 positions (left plot) and between the conserved asparagine ladder in repeat A and V_{3A} (right plot). The ten most commonly observed residue pairs are included for each pair of positions. (B) Frequencies of residue pairs in the conserved asparagine positions in LRR pairs for bacterial (left) and typical (right) LRR subfamilies. Bar charts are plotted as in (A) with all ten most common pairs (left subplots) and excluding the most common pair (right subplots).

The MI matrix also indicates significant inter-repeat coupling between the asparagine ladder positions (Figure 3.2; residues N_A and N_B, Figure 3.3A). Although the numerical value of the MI score between asparagine ladder positions is lower than that

between V_{3A} and V_{3B} (0.14 vs 0.22, Figure 3.2), the strong conservation of asparagine limits the MI score[a]. A closer look at the frequencies of residues found in each asparagine ladder position shows that the majority of LRR_BAC_PAIR sequences (84%) have two asparagines (Figure 3.3B). If residues at the two ladder positions assorted randomly, we would expect a frequency of 79% (the product of the probability of asparagine at position 12 and 32 in the multiple sequence alignment). Common substitutions include serine, threonine, and cysteine; of these residues, threonine is the only residue that occurs frequently in tandem (3.2% of all tandems, compared to an expectation value of 0.20%). The high frequency of tandem threonines in the bacterial LRR pairs is unique: the typical LRR subfamily does not show any appreciable frequency of tandem substitutions (Figure 3.3B).

Other positions that show strong correlations between repeats include positions of charge alternating substitutions between cLRR A and B [9]. The LRR_BAC_PAIR MI matrix has high scores for positions with high frequencies of charged residues, both within repeats (e.g. positions 6 and 10) and between repeats (e.g. position 6, 15, 16, and 19; Figure 3.2B). These positions correspond to charge alternating substitutions between the A and B cLRR repeats (residues 5, 10, 15, 16, and 19).

3.2.3 Nearest-neighbor modeling of cLRR constructs.

The charge alternating strategy employed in the cLRR sequence design improves the stability and solubility of cLRR constructs, but it also impacts nearest-neighbor

[a] For two positions, the mutual information value approaches zero as the positions become more conserved even though the residues are becoming more correlated (see equation 3.1).

modeling. For most studies that use nearest-neighbor models to describe unfolding of consensus repeat proteins, arrays with different numbers of identical repeats are analyzed to determine intrinsic stabilities of individual repeats and interfacial energies between repeats. When there are different repeat types, as with the A and B repeats of cLRRs, and each construct has the same number of A and B repeats, the intrinsic stabilities of individual repeats cannot be determined. Instead, a "paired repeat" model is required, which gives the stability of an AB repeat pair (R, Figure 3.1B) [9], along with the interfacial energy between repeat pairs (Figure 3.4, left). However, by including two additional constructs (NR_3A, BR_3C), the symmetry between A and B repeats is broken, allowing the stabilities of the individual repeats to be determined (Figure 3.4, right).

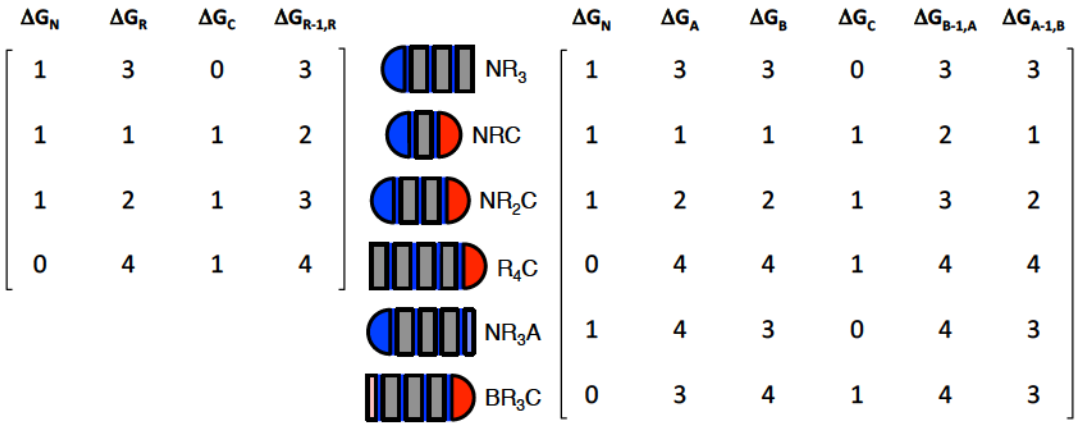


Figure 3.4. Matrices for paired-repeat and single repeat models. Comparison of matrices resolving intrinsic and interfacial parameters for the paired- (left) and single-repeat models (right). The cartoons represent constructs composed of N-cap (N, blue half circle), paired repeats (R, gray bar), and single repeats (A, thin powder blue bar; B, thin pink bar), and C-cap (C, red half circle). Column labels represent the intrinsic and interfacial terms resolved by each set of constructs. Both matrices are full rank.

Although NR_3A was soluble in the solution conditions used previously, BR_3C aggregated. We found that we could suppress this aggregation by adding 5% (v/v)

glycerol (Figure 3.5A), allowing us to measure reproducible equilibrium unfolding transitions. To account for the effects of glycerol on the stability of BR_3C, we measured unfolding transitions of NR_2 and NRC at two different glycerol concentrations (2.5%, 5% v/v). These were combined with unfolding transitions of all constructs in the absence of glycerol and the unfolding transition for BR_3C in 5% glycerol to globally fit using the single-repeat model, with both urea and glycerol dependences (Figure 3.5B) [8].

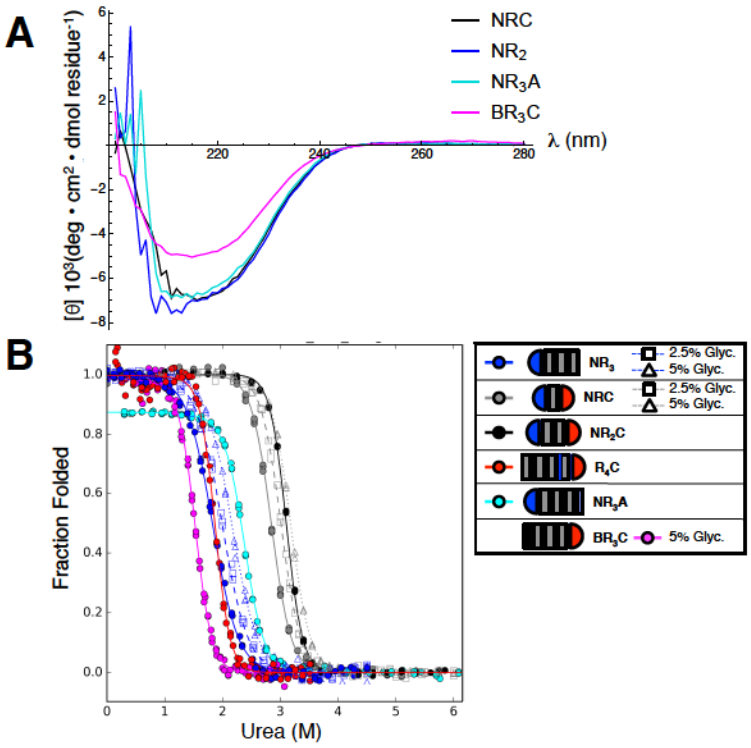


Figure 3.5. CD spectra and urea-induced unfolding of cLRR single-repeat model constructs. (A) Far-UV CD spectra of additional constructs required for single-repeat analysis. The BR_3C spectrum was taken in the presence of 5% glycerol (v/v). The signal intensity for BR_3C has a large uncertainty because association at high concentrations prevented accurate concentration determination. (B) Urea-induced unfolding transitions for cLRR constructs required to resolve single-repeat parameters. Data are fitted using a single-repeat nearest-neighbor model, and are normalized to give the fraction of folded repeats as a function of urea concentration Solid lines show fitted curves for each construct. Solution conditions: 20 mM $NaPO_4$, 500 mM NaCl, 0.1 mM TCEP, 0/2.5/5% glycerol (v/v), pH 7.8, 20 ˚C.

The joint urea/glycerol-dependent single-repeat model fits well to the data (solid curves, Figure 3.5B) with the surprising result that the NR_3A construct is partly unfolded in the absence of denaturant. As both NR_2 and NR_3 are fully folded in the model (NR_2, Figure 3.5B; NR_3, data not shown), the most likely explanation is that the terminal A repeat is unfolded in this construct. Though the far-UV CD spectrum of NR_3A is nearly identical to that of other fully folded constructs (Figure 3.5A), the weak contribution of β-sheet structure to CD signal and the relatively small size of a single repeat in the NR_3A construct (7 consensus repeats with 3 additional LRR repeats and an α-helical capping motif in the N-cap) may make far-UV CD insensitive to unfolding of the terminal A repeat. Consistent with this explanation, the fitted intrinsic folding free energy of the A repeat (ΔG_A) is larger than that of the B repeat (ΔG_B, Table 3.1), and the interfacial free energy for a B:A interface ($\Delta G_{B\text{-}1,\,A}$) is less stabilizing than that for an A:B interface ($\Delta G_{A\text{-}1,\,B}$, Table 3.1). As a result, the free energy of adding a single A repeat to the C-terminus of a folded cLRR array is positive (i.e., $\Delta G_A + \Delta G_{B\text{-}1,\,A} = 16.5 - 14.2 = 2.3$ kcal mol^{-1}).

Table 3.1. Fitted parameter values for paired- and single-repeat nearest neighbor parameters for cLRR unfolding.

Paired-repeat model		Single-repeat model	
ΔG_N	1.76 [1.68, 1.86]	ΔG_N	1.63 [1.52, 1.74]
ΔG_C	6.24 [6.08, 6.44]	ΔG_C	6.00 [5.75, 6.24]
ΔG_R	10.95 [10.75, 11.18]	ΔG_A[1]	16.52 [14.27, 19.51]
		ΔG_B	13.02 [12.73, 13.32]
$\Delta G_{R-1,R}$	-14.20 [-14.50, -13.95]	$\Delta G_{B-1,A}$	-14.15 [-14.47, -13.78]
		$\Delta G_{A-1,B}$[1]	-18.64 [-21.64, -16.42]
		$\Delta G_A + \Delta G_{A-1,B}$[2]	-2.13 [-2.24, -2.02]
m_{urea}	0.74 [0.75, 0.73]	m_{urea}	0.38 [0.37, 0.39]
m_{glyc}	0.36 [0.35, 0.38]	m_{glyc}	0.18 [0.18, 0.19]

Mean best-fit values from 100 (paired-repeat) or 2000 (single-repeat) bootstrap iterations. [1]Parameters could not be accurately determined because the NR_3A construct is partly unfolded (Figure S3.4). [2]The sum of the two undetermined parameters is well resolved. Confidence intervals for parameters are included in brackets. Free energies are in kcal mol^{-1}; m_{urea} and m_{glyc} values are in kcal mol^{-1} M_{urea}^{-1} and M_{glyc}^{-1}. Intrinsic folding free energies are represented as ΔG_i; interfacial free energies are represented as $\Delta G_{i-1,i}$, m_{urea} and m_{glyc} values give denaturant dependences for paired or single repeats. Denaturant dependences of N- and C-terminal caps are constrained to 1.5 m_{urea} and 3 m_{urea} in the paired-repeat model, and 3 m_{urea} and 6 m_{urea} in the single-repeat model to account for the number of repeats in each cap.

A consequence of the unfolding of the C-terminal A repeat in the NR_3A construct is that ΔG_A and $\Delta G_{A-1,B}$ are not well-determined by the six constructs analyzed here. This is illustrated in Figure S3.5, which shows the coefficient matrix for the six constructs with (A) and without (B) the terminal A repeat folded in NR_3B. Although the matrix has a rank of six when the terminal A repeat is folded, when it is unfolded the second and sixth column are identical, and thus the matrix has a rank of five. Thus, when the A repeat is unfolded, the coefficient matrix is not invertible, preventing solution for the six unknown free energy terms. Though the methods we use to solve for these coefficients (nonlinear

least squares using the full partition function) differs significantly from simple matrix inversion, it still suffers from the same drawbacks: our set of unfolding transitions lacks the information needed to determine all six free energies.

Although the full set of parameters cannot be recovered from the single-repeat model, tightly-constrained values are obtained for ΔG_B and $\Delta G_{B\text{-}1,A}$, as well as the urea dependence of the free energy of folding for individual repeats (m_i). The intrinsic folding free energy of a B-repeat is almost 2 kcal mol^{-1} higher than the folding energy of a repeat pair, and is significantly higher than that of any other consensus repeat measured in our lab (Figure 3.6A). The high cost of folding repeats is offset by favorable interfaces as indicated by $\Delta G_{B\text{-}1,A}$. The m_{urea} value for the cLRR repeat is less cooperative than expected given the size of the repeat [17], but this is a common trend observed for all consensus repeat proteins measured (Figure 3.6B).

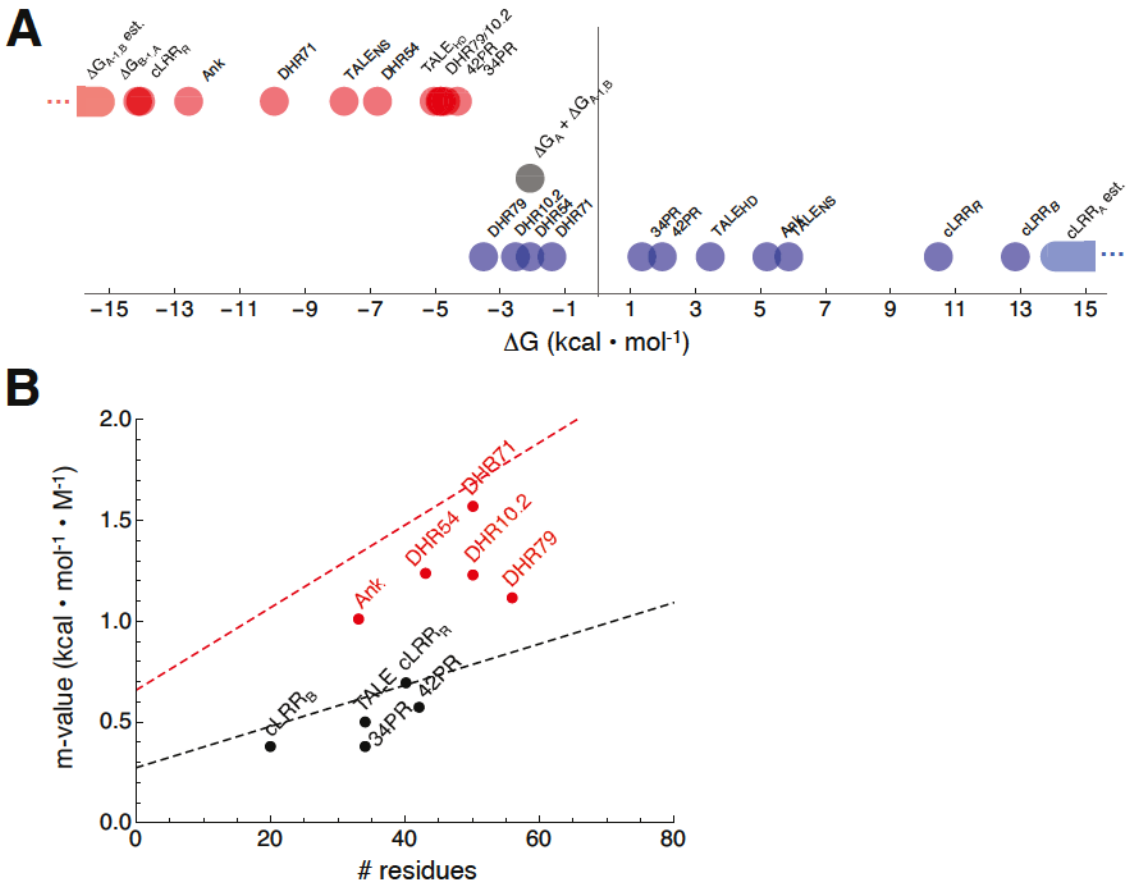


Figure 3.6. Comparison of cLRR single-repeat parameters to other consensus proteins. (A) ΔG values for intrinsic and interfacial parameters from studies of different consensus repeat proteins. Intrinsic (blue), interfacial (red), and the aggregated unresolved paired-repeat term (grey) are labeled by protein. Estimates for $\Delta G_{A\text{-}1,B}$ and $cLRR_A$ are shown as elongated points with an ellipsis to indicate that they are unbounded from below and above, respectively. (B) Comparison of consensus proteins m-values vs. their expected m-values based on their chain length [17]. Points represent the actual m-values of consensus proteins with the dashed line representing the relationship observed by [17] between ΔASA and urea m-value (black) and GdHCl (red). Denaturant used in consensus protein study is indicated by color (red, GdHCl; black, urea).

Though ΔG_A and $\Delta G_{A\text{-}1,B}$ cannot be individually determined by the model, their sum is well-determined from the fit. This is because these two parameters are tightly correlated (Figure S3.6A). Whereas distributions of bootstrapped parameters for ΔG_A and $\Delta G_{A\text{-}1,B}$ are very wide (with RMSD values of 1.7 kcal mol-1), the distribution of summed bootstrap values ΔG_A and $\Delta G_{A\text{-}1,B}$ is very tightly determined, with a standard deviation of 0.1 kcal

mol^{-1} (Figure S3.6B), and a mean value of -2.1 kcal mol^{-1}. This well-determined sum can be combined with the observation that the C-terminal A repeat in NR_3A is not folded to obtain limits on both ΔG_A and $\Delta G_{A\text{-}1,B}$. First, the observation that the C-terminal A repeat in NR_3A is not folded means that the sum of ΔG_A and $\Delta G_{B\text{-}1,A}$ greater than zero:

$$\Delta G_A + \Delta G_{B\text{-}1,A} > 0 \qquad 3.1$$

or

$$\Delta G_A > -\Delta G_{B\text{-}1,A} \qquad 3.2$$

This rearrangement expresses the idea that the stability of the B:A interface is insufficient to drive the folding of the A repeat. The value of $\Delta G_{B\text{-}1,A}$ is well determined (-14.2 kcal mol^{-1}), and can be combined with inequality 3.2 to provide a lower limit that $\Delta G_A > 14.2$ kcal mol^{-1}. This inequality can be combined with the sum $\Delta G_A + \Delta G_{A\text{-}1,B} =$ -2.1 kcal mol^{-1} to give an upper limit for $\Delta G_{A\text{-}1,B}$:

$$\Delta G_{A\text{-}1,B} - 2.1 = -\Delta G_A < -14.2 \text{ kcal mol}^{-1} \qquad 3.3$$

or

$$\Delta G_{A\text{-}1,B} < -16.3 \text{ kcal mol}^{-1} \qquad 3.4$$

(note that the inequality in equation 3.3 is reversed by the change of sign).

Comparing these inequalities to the well-determined parameters ΔG_B and $\Delta G_{A\text{-}1,B}$ shows that the A repeat is significantly less stable than the B repeat by at least 1.2 kcal mol^{-1}. The lower limit established here for ΔG_A makes the cLRR A repeat the least stable repeat measured in nearest-neighbor studies of repeat proteins to date (Figure 3.6).

Likewise, the upper limit for $\Delta G_{A\text{-}1,B}$ makes the A:B interface the most stable interface measured to date.

3.3 Discussion

3.3.1 Conservation and couplings between conserved cLRR positions.

The most significant difference between LRR_BAC_PAIR and those from the dataset used to construct the cLRR sequence [9] is the conservation pattern at position nine ($V_{3A/B}$, Figure 3.3A). This position is the third highly conserved leucine in the invariant LRR region (LxxLx***L***xxNxL). Although bulky hydrophobics (leucine, valine, isoleucine) make up a majority (54%) of the observed amino acids at this position in the LRR_BAC_PAIR alignment, there are a substantial number of cysteine-containing sequences (24%) that were not seen in the original alignment (Figure S3.1). Although misalignment of the LRR_BAC_PAIR sequences could be responsible for the surprising prevalence of cysteine at $V_{3A/B}$, cysteine is also observed (at lower frequencies) in the original dataset (Figure 3.3B). The presence of cysteine at $V_{3A/B}$ is particularly interesting given the significant destabilization of threonine substitutions at this position in the LRR protein pp32 ($\Delta\Delta G$ = 2.5 kcal mol^{-1}, see Chapter 2). The backbone CO of the $V_{3A/B}$ position is the canonical hydrogen bond acceptor for the asparagine ladder, but the thiol/hydroxyl side chains of cysteine/threonine conceivably may compete with their own backbone CO for the H_Z from the asparagine side chain (depending on the side chain rotamer). Alternatively, the preference for cysteine may derive from packing preference as it is bulkier than alanine but less costly to de-solvate than serine.

Another interesting result from the LRR_BAC_PAIR coupling data is the occurrence of tandem threonines at the ladder asparagine positions (Figure 3.3B) and cysteines at one of the conserved leucine positions (Figure 3.3A). These features do not seem to correlate with each other (i.e., cysteine at the $V_{3A/B}$ position is not frequently observed with threonine at the conserved asparagine position; Figure 3.3A, right panel) but do indicate that other ladders besides the commonly observed leucine, asparagine, and phenylalanine ladders may be present in the hydrophobic core of LRR proteins. It is likely that the threonine ladder leads to a reduction in stability and/or cooperativity, as substitutions away from the consensus asparagine result in less stable/cooperative variants of pp32 (see Chapter 2). The impact of the cysteine ladder is less clear, though substitutions to the conserved leucine positions in both pp32 and YopM also resulted in decreased stability [15], [18]. More detailed studies will be required to understand the effects of these deviations from the more common LRR sequence.

The MI patterns also indicate that the asparagine ladder is coupled, which has been experimentally determined via double-mutant cycles in pp32 (see Chapter 2). The MI plot seems to indicate weaker coupling than expected given the 5 kcal mol^{-1} coupling observed in pp32 but the highly conserved asparagine at these positions depresses the MI value. More variability can be observed at these positions if LRR pairs containing asparagines at both ladder positions are excluded from the analysis. Although there are not enough remaining sequences in LRR_BAC_PAIR to do an analysis of this kind, other LRR subfamilies can be used to show that MI values between the conserved asparagine

positions increase significantly after removing sequences with tandem asparagines (Figure S3.8).

The asparagine ladder position is also notable for its low coupling to other positions. Besides the relatively high coupling observed between ladder positions, the only other position with any appreciable MI is the following residue (position 13). Even removing LRR pairs with tandem asparagine from larger LRR sequence sets (as described above) does not generate appreciable MI between the asparagine ladder and other positions (Figure S3.8B). It is possible that loss of the conserved asparagine promotes stronger coupling between other positions to compensate for the instability resulting from ladder substitution. The MI profile for MSAs containing tandem asparagines (position N_A and N_B both asparagine, Figure 3.3A) or excluding tandem asparagines (either N_A or N_B can be asparagine but not both) can be compared to determine if coupling between other positions (e.g. better charge balancing) is strengthened. Subtracting the MI profiles from the tandem asparagine and non-tandem asparagine MSAs indicates that couplings increase among the first three conserved leucine positions in the invariant LRR region (***L***xx***L***xx***L***xxNxL), but only for the first repeat in the sequence (Figure S3.8B). It is odd that there are stronger MI values in only one repeat of the non-tandem asparagine dataset, which may be an artifact arising from the clustering of LRR pairs during the sequence analysis and so must be interpreted with caution. It is likely that decreased conservation is the cause of the higher MI values at L_1, L_2, and L_3. Since substitutions away from the invariant LRR sequence are destabilizing (see Chapter 2) [18], increased variability at these positions may reflect that the instability of the repeats is biologically

significant. A deeper analysis investigating the conservation patterns at work here and the use of more sophisticated detection methods (e. g. DCA [19]) should be brought to bear on this problem to identify subtle patterns that may be at work.

3.3.2 Stability of A versus B repeats in cLRR constructs.

The MI data also provided insights into the coevolution of polar/charged positions that might explain why A repeats are less stable than B repeats. Of the other charge alternating positions in the MI matrix, very few show any intra-repeat coupling and both A and B repeats have similar net charge magnitudes (-2 in A and +3 in B). The positions that do show large intra-repeat MI values are positions 6, 8, and 10; among these positions, A and B repeats differ only at position 10 (aspartate in A and serine in B, Figure 3.1B). The consensus sequence shows this position is most frequently a serine as in the B repeat (Figure S3.1A), which may explain the difference in intrinsic stability between the two repeats. The non-consensus phenylalanine at position eight may make the aspartate at position 10 in repeat A particularly unfavorable, as a bulky hydrophobic at position 8 with a negatively charged residue at position 10 occurs infrequently (0.5%) relative to what would be expected from random chance (1.3%).

The non-consensus phenylalanine at position eight may also contribute to the stability difference between B:A and A:B interfaces. A homology model of the cLRR sequence shows that the phenylalanine at position eight stacks with the phenylalanines from adjacent repeats, forming a phenylalanine stack (Figure 3.7A). Interestingly, phenylalanine ladders are commonly observed in the parallel β-sheet structure of amyloid

fibrils [20], where they contribute stabilizing π stacking interactions between adjacent repeats. As position eight has strong coupling to the residue at position ten, it is possible that interactions between the phenylalanine and aspartate (steric occlusion, competition) in repeat A disrupts the phenylalanine stacking. The homology model lends some support to this conclusion, as aspartates in repeat A are closer to the phenylalanine stack than the serines in repeat B (Figure 3.7B). However, this hypothesis must be tested by solving the cLRR structure and studying the effects of substitutions to the intra-repeat charge network (positions six, eight, and ten).

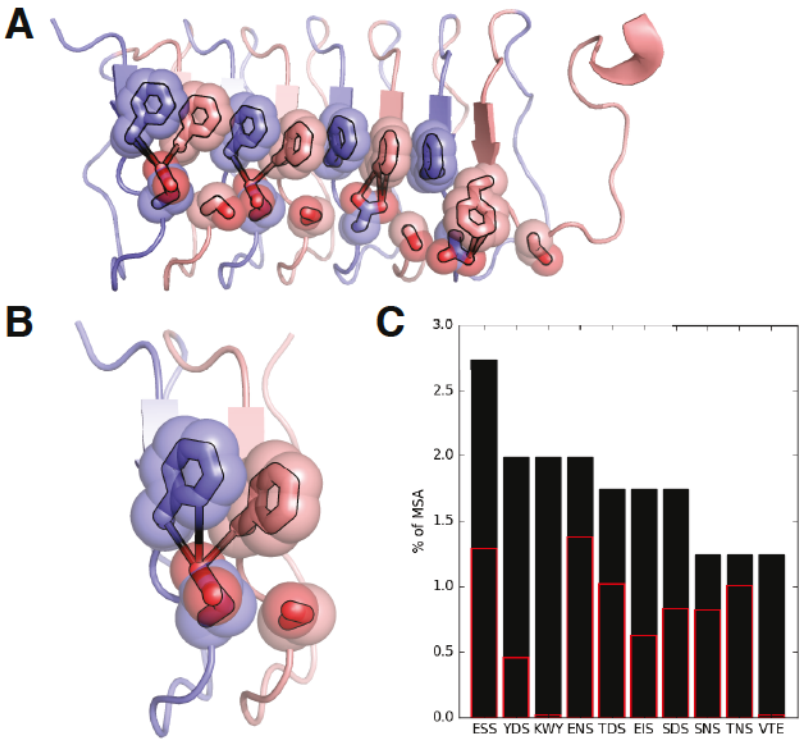


Figure 3.7. Homology models suggest a putative phenylalanine ladder. Homology model and repeats colored as in Figure 3.1A with positions eight (phenylalanine) and ten (repeat A, aspartate; repeat B, serine) shown with sticks and spheres. Distances between positions eight and ten within 4Å are highlighted with lines joining the atoms. **(A)** A model of four paired repeats without N- and C-terminal caps. **(B)** A single cLRR paired repeat highlighting potential impact of aspartate in repeat A. **(C)** Triplet frequencies for positions six, eight, and ten. Frequencies displayed as in Figure 3.3. Notably, all frequencies are very low, implying no single favored triplet permutation.

3.3.3 Trends in consensus protein nearest-neighbor model parameters.

cLRR repeats and interfaces are an extreme case of a general trend seen in most consensus proteins studied thus far. Most consensus proteins have repeats that are unstable when isolated (*de novo*-designed DHR proteins are the exception) but form highly favorable interfaces that support the unstable repeats. cLRRs follow this distribution with their intrinsic terms being the most unfavorable and interfacial terms the most favorable of any consensus protein studied thus far (Figure 3.6A). One interesting point of comparison is the TALE system, where Ising parameters for two different repeat types (HD and NS repeats) were determined [7]. Like the TALE system, cLRR A and B repeats have different interfacial and intrinsic free energies. Unlike TALEs, the cLRR A repeat remains unfolded in the absence of a C-terminal interface with a B repeat (Figure 3.5B). This behavior is reminiscent of the LRR protein YopM, where removing one interface by truncating the protein results in unfolding of multiple repeats [15]. Though the inability of repeat A to fold without an A:B interface is interesting, it would be worthwhile to redesign the repeat or explore new solution conditions permitting A to remain folded and the full set of single-repeat parameters to be determined.

Another trend observed from aggregation of multiple studies with consensus proteins is the linear anti-correlation of intrinsic and interfacial terms (Figure S3.7). It is notable that this trend holds not only for consensus proteins designed from naturally occurring sequences, but also for *de novo* designed sequences [21]. A linear fit of the data suggests that increasing the stability of a repeat by 1 kcal mol^{-1} decreases the stability of its interface by 0.6 kcal mol^{-1} (Figure S3.7). It is likely that the addition of new

protein families and more thorough investigation of the sequence space within families will better define the relationship between the intrinsic and interfacial terms.

3.4 Materials & Methods

3.4.1 LRR MSA construction

The SMART database breaks full LRR protein sequences into individual repeat fragments and divides these individual repeats into a number of canonical LRR subfamilies (typical, ribonuclease inhibitor, cysteine-containing, atypical). Unfortunately, the bacterial subfamily is not populated in SMART so it was reconstructed from the other LRR subfamilies. First, all LRR families were collected into a single file (LRR_TOT) of over 400,000 sequences, with typical lengths ranging from 20 – 28 residues (Figure 3.1). Using phylogenetic data from each sequence, LRR_TOT was then divided into bacterial and eukaryotic sequences. As expected from LRR classification studies, the bacterial set is enriched in 20 residue sequences compared to the eukaryotic set (Figure S3.1B) [22], [23]. The bacterial sequences were further filtered to include sequences of lengths 19 to 21 residues to build the LRR bacterial subfamily sequence set (LRR_BAC), which contained 2,837 entries.

The LRR_BAC sequences could be aligned as individual repeats but their short length (19-21 residues) makes accurate alignment difficult and cannot report on inter-repeat coupling that has been observed in previous studies [15] (see Chapter 2). To increase sequence length and improve alignments, we used the SMART database headers to determine the positions of the LRRs within their parent proteins, and combined

adjacent LRR sequences into repeat pairs. The LRR_BAC data set was then modified so that repeats separated by less than half a repeat (≤ 10 residues) were joined to form a pair. A new sequence header was applied to each pair to indicate the residue numbers of each single repeat. The resulting dataset (LRR_BAC_PAIR) could then be aligned and evaluated for couplings.

Though LRR pairs are aligned more easily than single repeats, they are still short compared to the average protein sequence length, making them a challenge to align accurately using standard parameters with alignment software like MAFFT [24]. Redundant sequences were removed prior to alignment using CDHit [25], [26] with sequences sharing 90% pairwise identity being clustered together; after clustering like sequences, 402 non-redundant sequences remained. Multiple sequence alignments (MSAs) for all LRR families were then generated using MAFFT with both gap penalty and extension penalty at their max value. Alignment quality was tested using the number of gaps in columns from the 11-residue invariant LRR region (LxxLxLxxNxL), as this region is a hallmark of all LRR sequences [27].

3.4.2 Conservation and mutual information.

Web logos of LRR_BAC_PAIR MSAs were obtained using the Skylign web server [28]. MI matrices were constructed using in-house Python code obtained from M. Sternke, modified to include average product correction [16]. The reported MI value for each matrix position is

$$MI_{APC} = D_{KL} - \frac{\overline{MI}_i \overline{MI}_j}{\overline{MI}} \quad (3.1)$$

where $D_{KL}(i,j)$ is the Kullback-Leibler divergence between the amino acid probability distributions of column *i* and *j* and MI_i is the mutual information from column i.

3.4.3 Protein cloning and expression.

cLRR constructs were cloned using PCR to generate a fragment of the paired-repeat DNA sequence. Primers were then used to generate the remaining 5' and 3' nucleotides and incorporate unique overhangs for an in-house version of Golden Gate cloning [29] developed by Dr. Kathryn Geiger-Schuller. Constructs were then cloned into pET24a+ vectors with a His_6 tag for affinity chromatography purification.

cLRR constructs were expressed by adding a single colony from a fresh (≤ 2 weeks old) transformation of BL21 cells to 1L of auto-induction media [30]. Flasks were shaken overnight at 37˚C for 14 – 18 hours, at which time cells were pelleted, resuspended in lysis buffer (20 mM $NaPO_4$, 500 mM NaCl, 25 mM Imidazole, 0.1 mM TCEP, pH 7.4) and one EDTA-free protease inhibitor tablet (Roche) per 50 mL lysis buffer, and lysed using either three freeze-thaw cycles in liquid nitrogen with lysozyme or sonification. After lysis, cells were incubated with DNAse and 1 mM $MgCl_2$ to digest genomic DNA and were centrifuged for 30 minutes at 30,000 x g. For constructs with N-caps, the clarified cell lysate was then poured over 5mL of Ni-NTA resin (Thermo Scientific), washed in 50-100 mL of lysis buffer, and eluted in 50 mL of elution buffer (lysis buffer with 250 mM

imidazole). Constructs were then placed in 4L of dialysis buffer (20 mM $NaPO_4$, 500 mM NaCl, 0.1 mM TCEP, pH 7.8) for at least 24 hours, concentrated, and stored at -80 ˚C.

ΔN-cap constructs were expressed in inclusion bodies, so clarified cell lysate was removed and the pellets were resuspended in lysis buffer with 4 M urea (VWR Life sciences) and then spun again at 40,000 x g for one hour. The soluble portion of the resuspended pellet was then poured over the Ni-NTA resin and washed in 50 – 100 mL of the urea-fortified lysis buffer, followed by 50 mL of urea-fortified elution buffer. Constructs were then placed in 4 L of dialysis buffer with 10% glycerol (v/v) for at least 24 hours. ΔN-cap constructs were concentrated to ~30 μM to avoid concentration-dependent aggregation and stored at -80 ˚C.

3.4.4 Circular dichroism spectra and equilibrium unfolding.

All CD experiments were performed on an Aviv Model 400 CD spectrometer using a computer-controlled Microlab syringe titrator (Hamilton) with samples in CD buffer (20 mM NaPO4, 500 mM NaCl, 0.1 mM TCEP, 0/5% glycerol (v/v), pH 7.8) at 20 ˚C. CD spectra were collected with 3 second signal averaging every nm from 280 to 200 nm; protein concentrations were 30 to 50 μM in a 0.1 mm path-length quartz cuvette. Equilibrium unfolding experiments were monitored at 220 nm using a 5-6 min equilibration time and 30 s signal averaging; protein concentrations were 3-5 μM in a 1 cm path length quartz cuvette. Urea (VWR Life Sciences) used for denaturation studies was deionized using a mixed-bed resin (BioRad) immediately prior to use. Urea concentrations were determined by refractometry (Pace, 1986). NR_2 and NRC constructs were denatured in

0, 2.5, and 5% glycerol (v/v). BR_3C constructs were denatured in 5% glycerol (v/v). BR_3C unfolding experiments were performed titrating from high urea concentrations to lower concentrations with no protein in the titrant to avoid aggregation of titrant protein at low urea concentrations over the course of the experiment.

3.4.5 Nearest-neighbor analysis.

Determination of single-repeat interfacial and intrinsic parameters was achieved using a 1D Ising model [31]. Parameters were defined within equilibrium constants that quantify intrinsic folding and repeat-repeat interfaces,

$$\kappa_N = e^{-(\Delta G_N - m_i * x - m_g * g)\beta} \quad (3.2)$$

$$\kappa_A = e^{-(\Delta G_A - m_i * x - m_g * g)\beta} \quad (3.3)$$

$$\kappa_B = e^{-(\Delta G_B - m_i * x - m_g * g)\beta} \quad (3.4)$$

$$\kappa_C = e^{-(\Delta G_C - m_i * x - m_g * g)\beta} \quad (3.5)$$

$$\tau_{B-1,A} = e^{-\Delta G_{B-1,A}\beta} \quad (3.6)$$

$$\tau_{A-1,B} = e^{-\Delta G_{A-1,B}\beta} \quad (3.7)$$

where β is $(kT)^{-1}$, x is molar urea concentration, and g is molar glycerol concentration. Equilibrium constants for the N:A interfaces and B:C interfaces are simply set to that for the B:A interface ($\tau_{B\text{-}1,A}$) as any differences between $\tau_{B\text{-}1,A}$ and $\tau_{N\text{-}1,A/B\text{-}1,C}$ can be compensated for within the κ_N and κ_C terms and has been assumed previously [6]. The m_{urea}/m_{glyc} values for N- and C-caps are scaled to represent the number of repeats within each cap. Using these equilibrium constants, a partition function can be constructed and

used to determine the fraction of repeats folded in the array [31]. Parameters were determined using nonlinear least squares to globally fit normalized urea denaturation data [8] using a script designed by Dr. Jake Marold, modified by Dr. Katie Geiger-Schuller, and with additional minor modifications.

3.5 Supplementary Figures

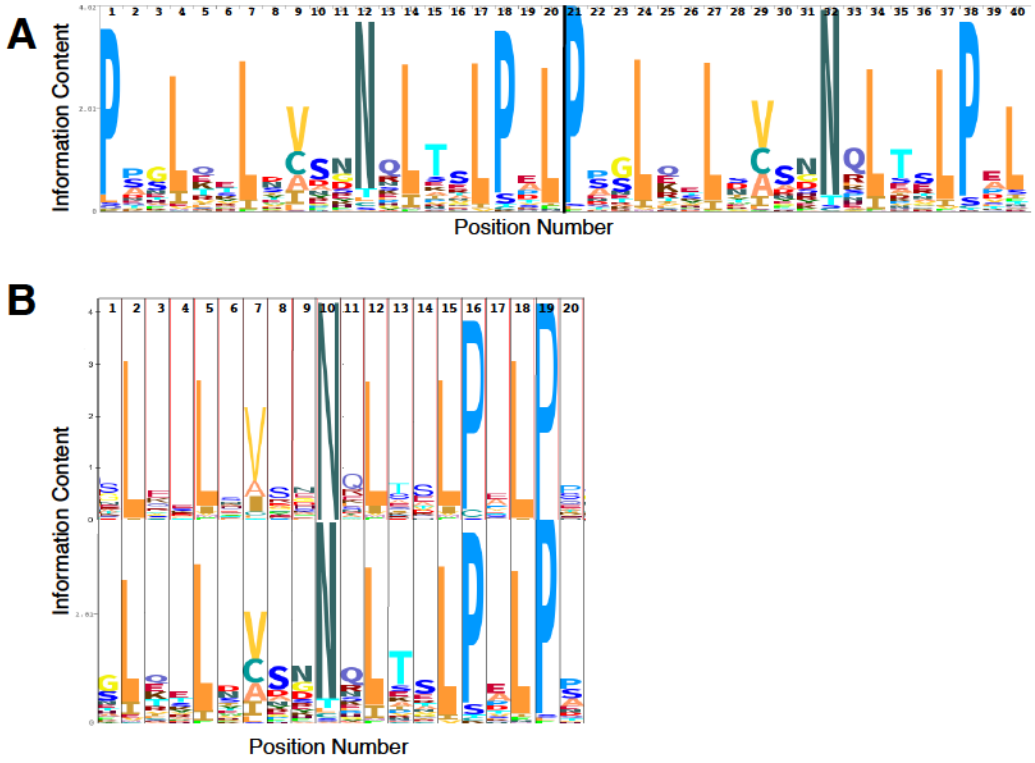


Figure S3.1. Conservation in bacterial LRR subfamily. (A) HMM logo of LRR_BAC_PAIR sequences representing the bacterial LRR subfamily. Position numbers are indicated above each column with a black line dividing the repeats. (B) HMM logos from the original consensus design (top) and a single repeat from the LRR_BAC_PAIR sequence set (bottom). Y-axis units are not the same but the conservation patterns are clear.

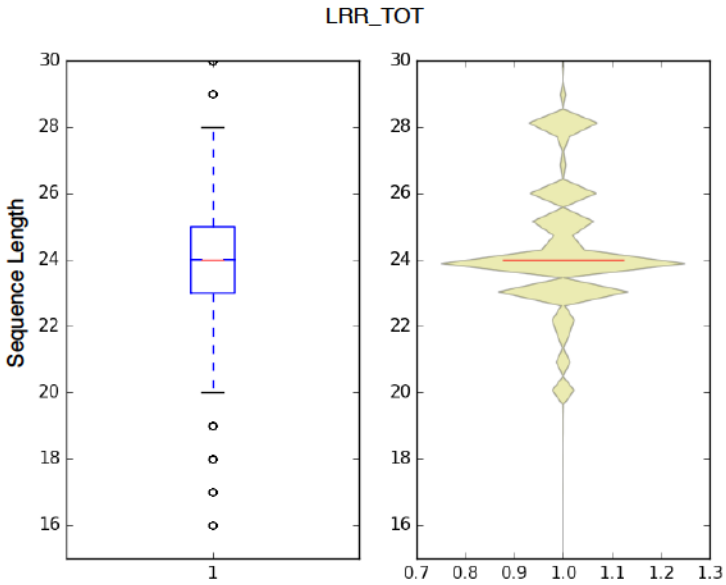


Figure S3.2. Sequence length distribution in LRR sequences. Boxplot (left) and violin plot (right) of the sequence length distributions in the LRR_TOT sequence set. (Left) Median and mean of the set are red and blue lines respectively. The blue box represents one standard deviation above and below the mean with the dashed lines extending to the first and fourth quartile of the data set. Outliers are shown as black circles. (Right) A solid red line represents medians of each subset. Sequence lengths above 30 and below 15 are excluded as these fall outside the lengths of known LRR subfamilies.

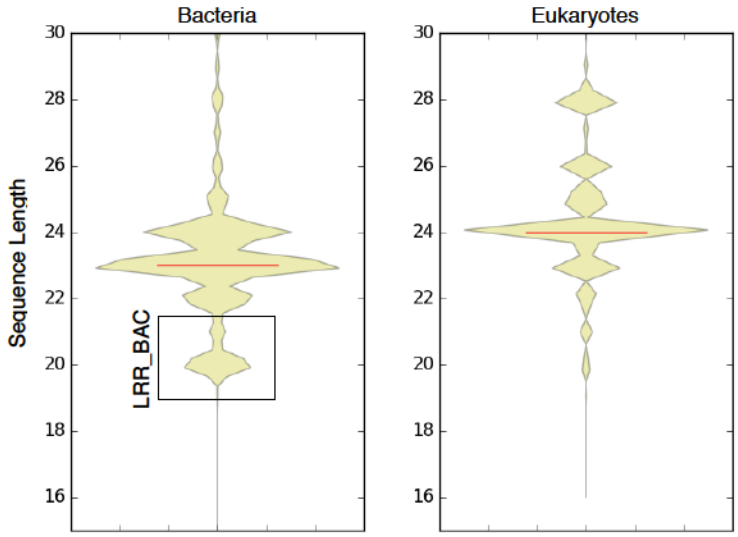


Figure S3.3. Sequence length distribution in LRR subfamilies. Comparison of the sequence length distributions for the bacterial and eukaryotic subsets of the LRR_TOT sequence set. Sequence lengths above 30 and below 15 are excluded as these fall outside the lengths of known LRR subfamilies. Medians of each subset are represented by a solid red line. The black rectangle on the bacteria subset indicates the sequences used to represent the bacterial

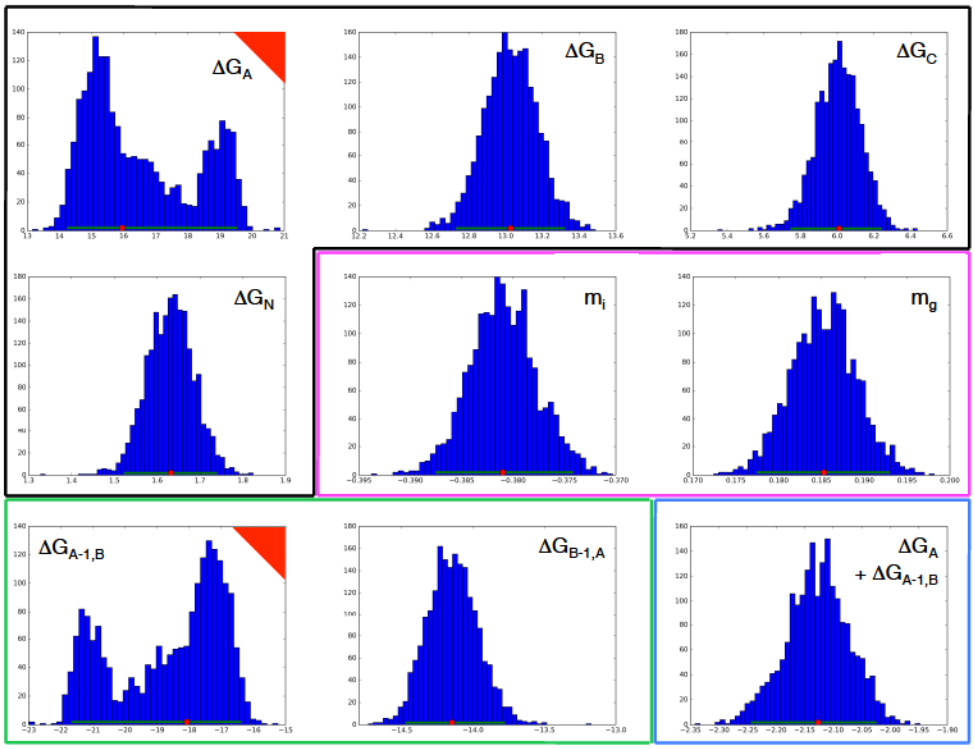


Figure S3.4 Uncertainty in parameters from single-repeat model of cLRR. Parameter value distributions from 2000 bootstrap iterations of microscopic nearest-neighbor model. Intrinsic (black box), urea and glycerol dependence (magenta box), and the original interfacial terms (green box) parameter values are shown as histograms with the parameter name in the top right of the subplot. As ΔG_A and $\Delta G_{A\text{-}1,B}$ could not be uniquely determined, the values for ΔG_A and $\Delta G_{A\text{-}1,B}$ were summed for each fit and called as a single parameter (ΔG_A + $\Delta G_{A\text{-}1,B}$, blue box) that was well-determined. Mean value is demarcated by a red point with 95% confidence intervals extending in green through this point. Units for subplots are kcal mol^{-1} (intrinsic, black; interfacial, green; dG_A + $dG_{A\text{-}1,B}$, blue) and kcal mol^{-1} $M_{urea/glycerol}$ (mi/mg, magenta). Undetermined parameters are demarcated by a red triangle in the upper right of the subplot.

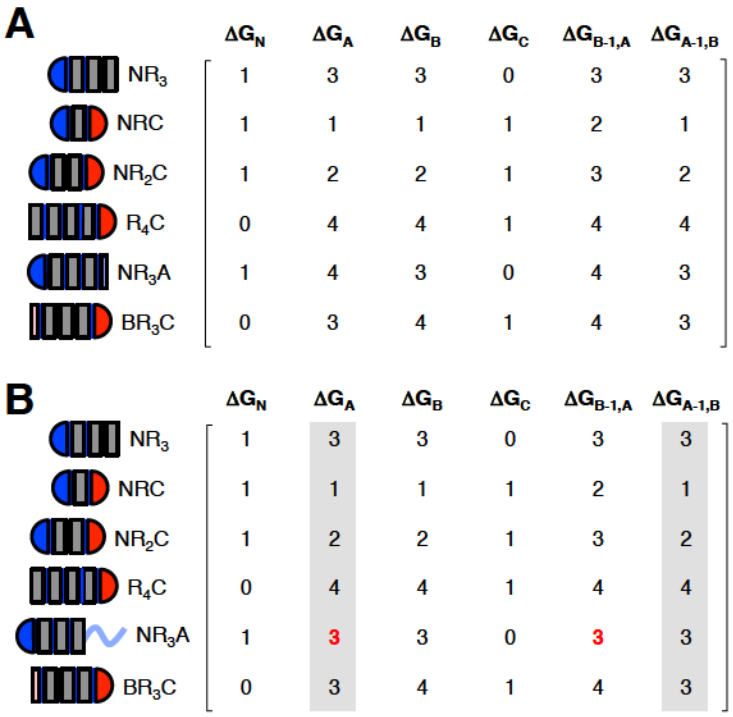


Figure S3.5. Comparison of single-repeat matrices for all constructs fully folded and N-terminal A unfolded. Cartoons and matrices are represented as in Figure 3.4 with the unfolded terminal A repeat in the NR_3A construct of (B) shown as a powder blue line. (A) With all constructs fully folded, the matrix is full rank and all parameters can be determined. (B) When the terminal A repeat is unfolded, the columns associated with the ΔG_A and $\Delta G_{A\text{-}1,B}$ terms (gray boxes) are identical and unresolved, reducing the rank of the matrix to five. The NR_3A matrix positions changed by the unfolded A repeat are in red.

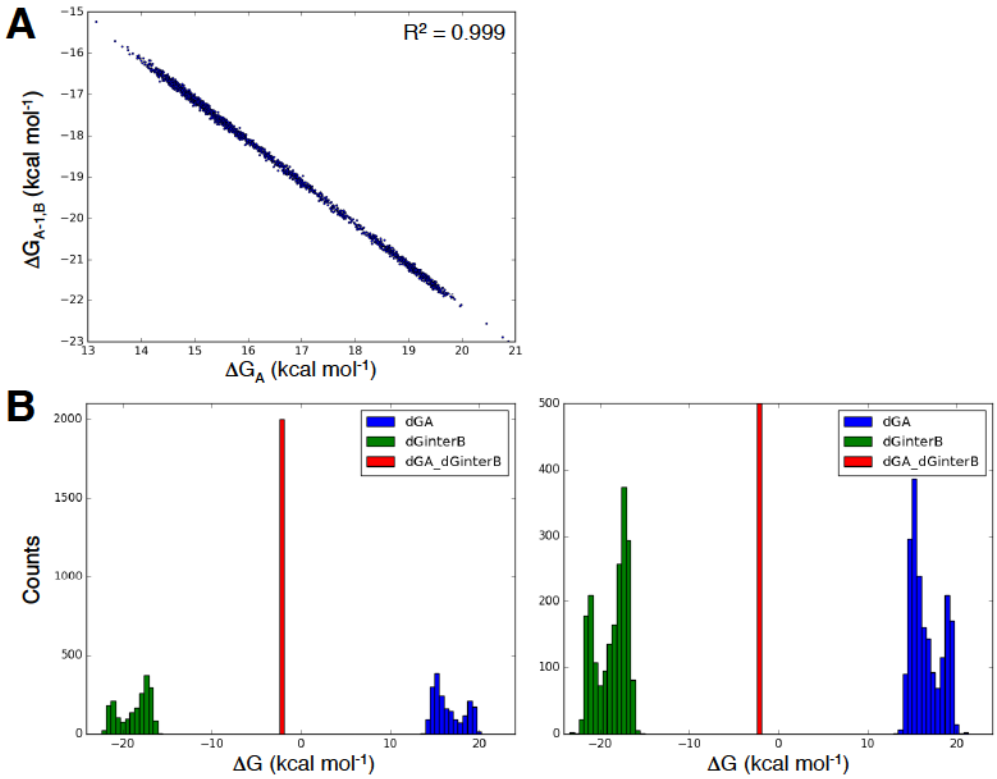


Figure S3.6. Correlation and uncertainty in ΔG_A and $\Delta G_{A\text{-}1,B}$ parameters from single-repeat model. (A) Correlation between ΔG_A and $\Delta G_{A\text{-}1,B}$ from 2000 fitting iterations. R^2 value at top right corresponds to a linear fit of the data. (B) Comparison of uncertainties for ΔG_A (blue), $\Delta G_{A\text{-}1,B}$ (green), and their sum (red). Left plot shows the full range of y-axis values and the right plot shows only counts below 300 to better demonstrate the uncertainty in ΔG_A and $\Delta G_{A\text{-}1,B}$.

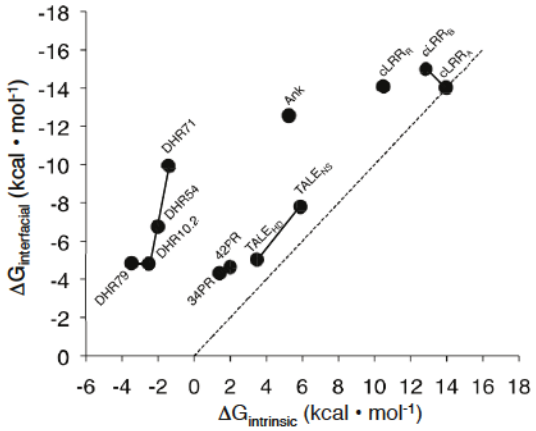


Figure S3.7. Correlation between interfacial and intrinsic ΔG for consensus proteins. Proteins from similar families are joined by line. The dashed line represents a slope of one. A linear fit of the data has slope -0.6 and $R^2 = 0.656$.

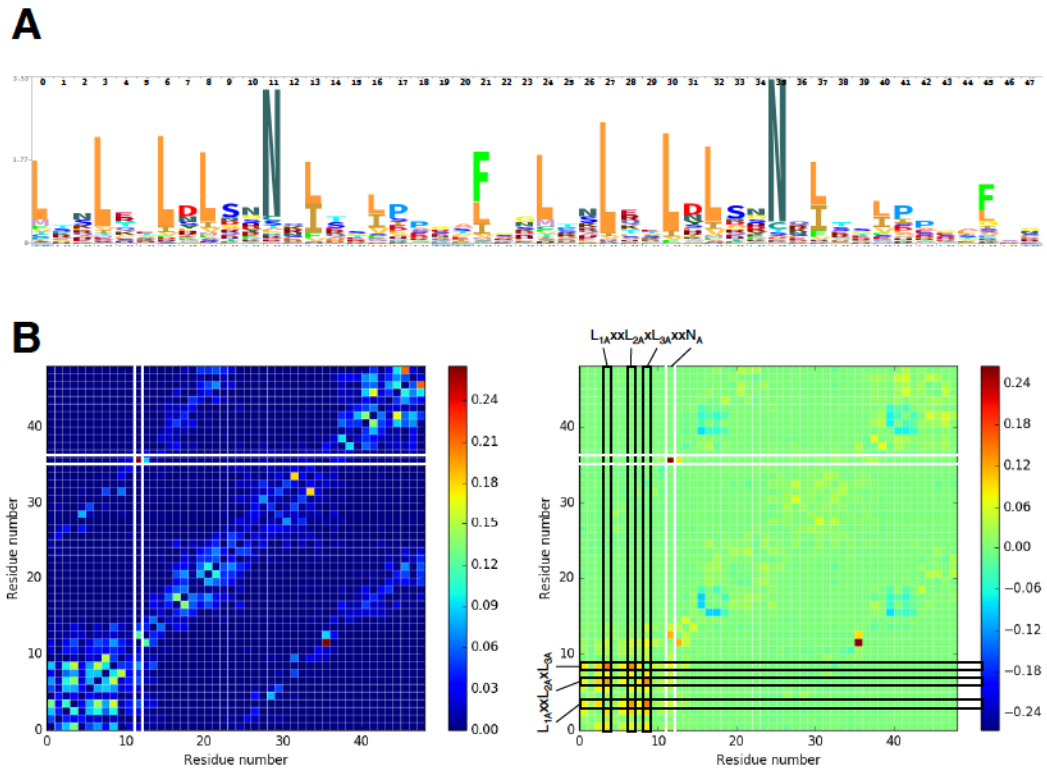


Figure S3.8. Analysis of couplings from LRR typical subfamily sequences. (A) HMM logo from LRR typical sequences with position numbers corresponding to matrix positions in (B). (B) MI matrix for LRR typical sequences with no more than one asparagine at the N_A and N_B positions (labeling scheme in Figure 3.1B, left) and the MI matrix resulting from $MI_{non\text{-}tandem}$ - MI_{tandem} (right). Conserved asparagine positions are highlighted by white rectangles and the first three conserved leucine positions from the LRR invariant sequence are highlighted by black rectangles. MI value scales are located to the right of each matrix.

3.6 References

[1] E. Kloss, N. Courtemanche, and D. Barrick, “Repeat-protein folding: New insights into origins of cooperativity, stability, and topology,” *Arch. Biochem. Biophys.*, vol. 469, no. 1, pp. 83–99, 2008.

[2] Z.-Y. Kan, B. T. Walters, L. Mayne, and S. W. Englander, “Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis,” *Proc. Natl. Acad. Sci.*, 2013.

[3] K. W. Tripp and D. Barrick, “Enhancing the Stability and Folding Rate of a Repeat Protein through the Addition of Consensus Repeats,” *J. Mol. Biol.*, vol. 365, no. 4, pp. 1187–1200, 2007.

[4] T. Kajander, A. L. Cortajarena, E. R. G. Main, S. G. J. Mochrie, and L. Regan, “A new folding paradigm for repeat proteins,” *J. Am. Chem. Soc.*, 2005.

[5] T. Aksel, A. Majumdar, and D. Barrick, “The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding,” *Structure*, vol. 19, no. 3, pp. 349–360, 2011.

[6] J. D. Marold, J. M. Kavran, G. D. Bowman, and D. Barrick, “A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins,” *Structure*, 2015.

[7] K. Geiger-Schuller and D. Barrick, “Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States,” *Biophys. J.*, vol. 111, no. 11, pp. 2395–2403, 2016.

[8] K. Geiger-Schuller, K. Sforza, M. Yuhas, F. Parmeggiani, D. Baker, and D. Barrick, “Extreme stability in de novo-designed repeat arrays is determined by unusually stable short-range interactions,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 29, p. 201800283, 2018.

[9] T. P. Dao, *Determination of folding pathway selection and origins of cooperativity of naturally occurring and designed consensus leucine-rich repeat proteins*. Johns Hopkins University, 2014.

[10] T. Aksel and D. Barrick, “Direct observation of parallel folding pathways revealed using a symmetric repeat protein system,” *Biophys. J.*, vol. 107, no. 1, pp. 220–232, 2014.

[11] M. R. Preimesberger *et al.*, “Direct NMR detection of bifurcated hydrogen bonding in the α-helix N-caps of ankyrin repeat proteins,” *J. Am. Chem. Soc.*, 2015.

[12] A. Waterhouse *et al.*, “SWISS-MODEL: Homology modelling of protein structures and complexes,” *Nucleic Acids Res.*, 2018.

[13] S. El-Gebali *et al.*, “The Pfam protein families database in 2019,” *Nucleic Acids Res.*, 2019.

[14] I. Letunic and P. Bork, “20 years of the SMART protein domain annotation resource,” *Nucleic Acids Res.*, 2018.

[15] E. Kloss and D. Barrick, “C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein,” *Protein Sci.*, vol. 18, no. 9, pp. 1948–1960, 2009.

[16] S. D. Dunn, L. M. Wahl, and G. B. Gloor, “Mutual information without the influence

of phylogeny or entropy dramatically improves residue contact prediction," *Bioinformatics*, 2008.
[17] J. K. Myers, C. Nick Pace, and J. Martin Scholtz, "Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding," *Protein Sci.*, vol. 4, no. 10, pp. 2138–2148, 1995.
[18] T. P. Dao, A. Majumdar, and D. Barrick, "Highly polarized C-terminal transition state of the leucine-rich repeat domain of PP32 is governed by local stability," *Proc. Natl. Acad. Sci.*, 2015.
[19] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, "Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 2013.
[20] J. Greenwald and R. Riek, "Biology of amyloid: Structure, function, and regulation," *Structure*, vol. 18, no. 10, pp. 1244–1260, 2010.
[21] T. J. Brunette *et al.*, "Exploring the repeat protein universe through computational protein design," *Nature*, vol. 528, no. 7583, pp. 580–584, 2015.
[22] A. V. Kajava, "Structural diversity of leucine-rich repeat proteins," *J. Mol. Biol.*, 1998.
[23] S. G. S. T. C. Buchanan and N. J. Gay, "Structural and functional diversity in the leucine rich repeat family of proteins.," *Prog. Biophys. Molec. Biol.*, vol. 65, no. I, pp. 1–44, 1996.
[24] K. Katoh, J. Rozewicki, and K. D. Yamada, "MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization," *Brief. Bioinform.*, 2017.
[25] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, 2012.
[26] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, 2006.
[27] B. Kobe and A. V. Kajava, "The leucine-rich repeat as a protein recognition motif," *Current Opinion in Structural Biology*. 2001.
[28] T. J. Wheeler, J. Clements, and R. D. Finn, "Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models," *BMC Bioinformatics*, 2014.
[29] T. Cermak *et al.*, "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting," *Nucleic Acids Res.*, 2011.
[30] F. W. Studier, "Protein production by auto-induction in high density shaking cultures.," *Protein Expr. Purif.*, 2005.
[31] T. Aksel and D. Barrick, *Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models*, 1st ed., vol. 455, no. A. Elsevier Inc., 2009.

INTENDED TO BE BLANK

CHAPTER 4 – Evaluation of asparagine ladder substitutions in a consensus leucine-rich repeat protein.

4.1 Introduction

Predicting protein structure [1], designing new or modified versions of proteins [2], and modulating protein function [3] are all challenging endeavors in the field of protein biophysics, structural biology, and evolutionary biochemistry. Part of the difficulty associated with these efforts stems from an inability to accurately predict higher order phenomena that arise from the network of interactions within and/or between proteins. A better understanding of how groups of connected residues interact to produce properties such as cooperativity in protein folding [4] would increase our ability to design novel structures, understand the molecular origins of disease, and develop treatments.

Understanding the interactions between residues in a protein requires perturbation through substitutions, deletions, or insertions. The response of the protein can then be measured, using the free energy of folding, binding, or some other metric to assay the effects of the perturbations. Unfortunately, a detailed understanding of the complex web of interactions in proteins requires the generation of a large number of variants [5], [6]. For most such studies, the networks of interest are composed of residues with widely varying physical properties and equally diverse interactions. More homogenous networks limit this variety, thereby reducing the number of variants needed to fully capture the behavior of the networked residues (see Chapter 2) [7].

Repeat proteins are particularly useful for studies of uncovering interaction networks, both for their structural simplicity and their compatibility with nearest-neighbor

model analysis. Repeat proteins are composed of linear arrays of a single motif, which reduces long-range contacts and local differences between the motifs; these properties make repeat proteins a good system for rigorous quantification of protein physical properties [8]. Furthermore, a consensus sequence for the repeated domain can be duplicated to generate folded proteins amenable to nearest-neighbor modeling [9], [10]. Using a nearest-neighbor model, perturbations to a homogenous network can be partitioned into contributions to coupling between repeats and folding of the repeats in isolation [11]. Therefore, a nearest-neighbor model's description of substitution effects would be more complete than the description provided by a traditional mutagenesis study.

An excellent system in which to study a homogenous network of interacting residues is the leucine-rich repeat (LRR) protein family. LRR proteins contain a highly conserved network of asparagine residues (asparagine ladder) embedded in their hydrophobic core (Figure 4.1) [12]. Studies in chapter 2 show that the asparagine ladder contributes to global stability and cooperativity but were not able to quantitatively resolve whether the ladder contributes to local stability or coupling between repeats. In this chapter, I use a nearest neighbor model with asparagine ladder substitutions to demonstrate that the asparagine ladder provides strong energetic coupling between repeats. I also determine that ladder substitutions perturb non-adjacent interfaces, especially interfaces that are N-terminal to the site of the substitution. I compare these findings to substitution studies in other consensus repeat proteins [13], [14] to determine if a general relationship between intrinsic and interfacial stability exists.

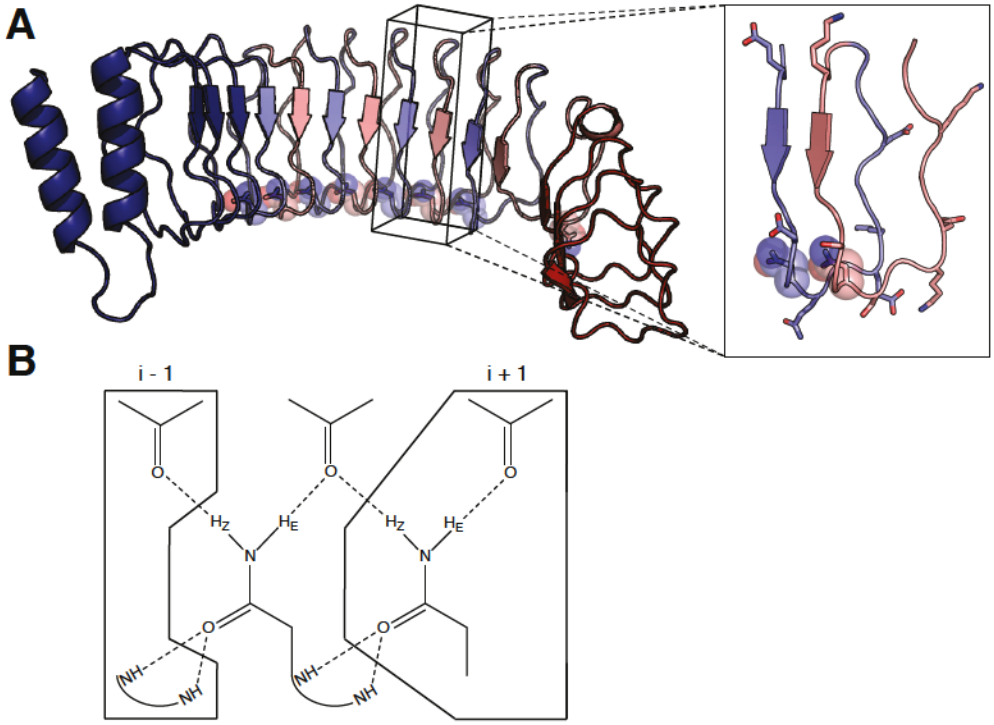


Figure 4.1. Structure of consensus LRR protein and asparagine ladder. (A) Homology models of cLRR NR_4C structure from SWISSMODEL webserver. The model is composed of an N-cap (dark blue), four paired repeats, each with a single A (light blue) and B (light red) repeat, and a C-cap (dark red). The putative asparagine ladder is shown in sticks and spheres. The inset shows a single AB repeat pair with charge swap and polar substitutions shown in sticks and the asparagine ladder position shown in sticks and spheres. (B) Scheme of molecular interactions for asparagine ladder at repeat *i*. Dashed lines represent hydrogen bonds and NHs are backbone amides. Contacts are shown between the ladder asparagine in repeat *i* and repeat $i \pm 1$ (in black polygons).

4.2 Results

4.2.1 Selecting substitutions from LRR conservation patterns.

The substitutions for this cLRR study were chosen from residues observed frequently and infrequently in an MSA of LRR sequences (Figure 4.2A). Cysteine was selected because it is the most frequently observed alternative to asparagine at the ladder position. Leucine was chosen to measure the effects of hydrophobic replacement of the

polar asparagine. Frequently observed residues should be less destabilizing than infrequent ones unless there are compensating substitutions required to stabilize them. Such compensating substitutions should show positive covariance with the infrequent ladder residue, which should be reflected in a mutual information statistic [15]. However, there is almost no mutual information between the conserved asparagine position and other positions within/between LRR repeats (see Chapter 3). Therefore, probable substitutions are expected to be more stabilizing than improbable ones at the asparagine ladder position with minimal dependence on the sequence context.

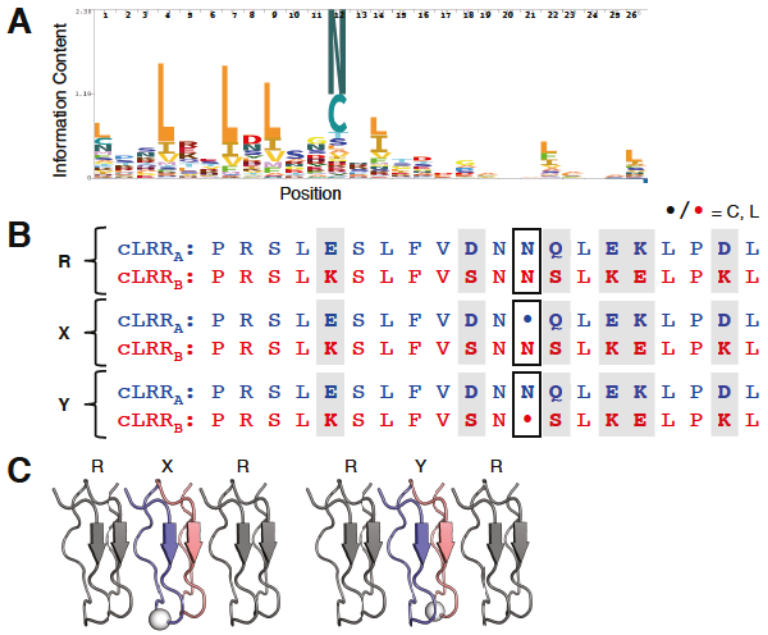


Figure 4.2. cLRR sequence conservation and substitutions. (A) Web logo for a collection of sequences from all LRR classes. Position numbers for the weblogo and the cLRR sequence in (B) are equivalent up to position 12 (the invariant LRR region), with the asparagine ladder residue occurring at position 10. (B) Sequences for consensus LRR (cLRR) protein. Each repeat pair (R) is composed of an A (blue) and B (red) repeat. Charge swap positions (filled gray rectangles) and the conserved asparagine position (empty black rectangle) are indicated. Substitutions can occur in repeat A (X) or B (Y) and are represented by a dot. (C) Sites of substitution within $X_{Cys/Leu}$ and $Y_{Cys/Leu}$ repeats. Repeats within $X_{Cys/Leu}$ and $Y_{Cys/Leu}$ repeat pairs are colored (A, blue; B, red) with unsubstituted repeat pairs shown in gray. Substitution sites are shown with a white sphere to emphasize adjacent and non-adjacent interfaces.

As cysteine is commonly observed at the asparagine ladder position in LRRs generally, it was treated as a probable substitution while leucine was treated as improbable (Figure 4.2A). However, sequence analysis of the bacterial LRR class, from which the original cLRR sequence was designed, indicates that cysteine is no more frequent than leucine (Figure S4.1). Although cysteine frequencies are low, the lack of mutual information at the asparagine ladder position suggests that there should not be compensating substitutions in the bacterial LRRs that might make cysteine more destabilizing than in other LRR classes (e.g., typical, cysteine-containing [16]). Therefore, cysteine should behave in the cLRR sequence much as it would in a sequence designed from any other LRR class.

4.2.2 Paired repeats model and constructs for cLRR substitutions.

Quantification of the contribution of the asparagine ladder to interfacial and intrinsic ΔG° values requires substitution of ladder asparagines in the cLRR sequence. Substitutions were chosen based on the ladder residue frequencies across all LRR classes, where cysteine is frequently observed and leucine is rare (Figure 4.2A). Given the asparagine ladder side chains make more contacts with *i-1*th repeat, it is important to resolve the preceding (N-terminal) and succeeding (C-terminal) interfaces into separate parameters. In addition, the AB repeat pair can be altered in two potentially distinct ways: substitutions can be incorporated into either the A repeat or the B repeat. To distinguish these two substitutions, we refer to a repeat pair with an A-repeat substitution as an "X" paired repeat (compared to a wild-type "R" paired repeat; Figure 4.2B, middle), and to a

repeat pair with a B-repeat substitution as a "Y" paired repeat (Figure 4.2B, bottom). Sequence differences between A and B result in different intrinsic stabilities (see Chapter 3) so it is expected that the effects of substitution on intrinsic stability will depend on whether the substitution is made to the A (X) or B (Y) repeats (Figure 4.2C).

As with the original cLRR sequence, these substitutions can be studied using either the paired-repeat (Figure 4.3) or single-repeat model. For this study, the paired repeats scheme is used. The residue substituted for the ladder asparagine will be indicated with a subscript, e.g, X_{Cys} and Y_{Leu}. A cursory look at the cLRR architecture shows that X_{aa} constructs will directly impact the N-terminal interface adjacent to the substitution site and Y_{aa} constructs will similarly impact the C-terminal interface (Figure 4.2C). However, the interface C-terminal to an X_{aa} substitution may also be affected through long-range interactions; likewise, the interface N-terminal to a Y_{aa} substitution may be affected. In both cases, these distal effects involve the $i+2^{th}$ and $i-2^{th}$ interface, respectively. Since these non-adjacent interfaces are resolved in the paired repeats model, they will be a useful measure of long-range coupling within the asparagine ladder.

$\Delta G_{X/Y}$	$\Delta G_{R-1,X/Y}$	$\Delta G_{X/Y-1,R}$	$\Delta G_{X/Y-1,X/Y}$	
1	1	0	0	 $NR_2(X/Y)$
1	1	1	0	 NR(X/Y)R
1	0	1	0	 $(X/Y)R_3C$
2	1	1	1	 $NR(X_2/Y_2)$

Figure 4.3. Matrix for resolving paired repeats parameters for both $X_{Cys/Leu}$ and $Y_{Cys/Leu}$ substitutions. The cartoons represent constructs composed of combinations of N-cap (N, blue half circle), R repeat pairs (gray bar), $X_{Cys/Leu}/Y_{Cys/Leu}$ repeat pairs (white), and C-cap (C, red half circle). Interfaces between un-substituted repeats are shaded blue, R:(X/Y) interfaces are shaded black, (X/Y):R interfaces are shaded red, and X:X or Y:Y interfaces are shaded gray. Column labels represent the intrinsic and interfacial terms resolved by the constructs. As long as all of the repeats are folded in the absence of denaturant, the matrix if full rank, and parameters can be accurately determined from urea-induced unfolding transitions of the four constructs shown.

Proteins matching the constructs from the matrix in Figure 4.3 with the ladder asparagine substituted for a cysteine or leucine in the A or B repeats (Figure 4.2B) were expressed and purified. As with the original cLRR constructs (see Chapter 3), ΔN-cap $X_{Cys/Leu}/Y_{Cys/Leu}$ variants required the addition of 5% glycerol (v/v) for solubility at low urea concentrations. $X_{Cys/Leu}/Y_{Cys/Leu}$-containing constructs had far-UV CD profiles similar to the original cLRR constructs (Figure 4.4A) implying they maintain a similar fold. Compared to the original cLRR sequence, internal and double substitutions ($R(X_{Cys/Leu}/Y_{Cys/Leu})R$, $R(X_{Cys/Leu})_2/(Y_{Cys/Leu})_2$) generally have a reduced CD signal near 210-220 nm. In previous studies, reduced CD signal was correlated with decreased stability and (in some cases) cooperativity (see Chapter 2) [17].

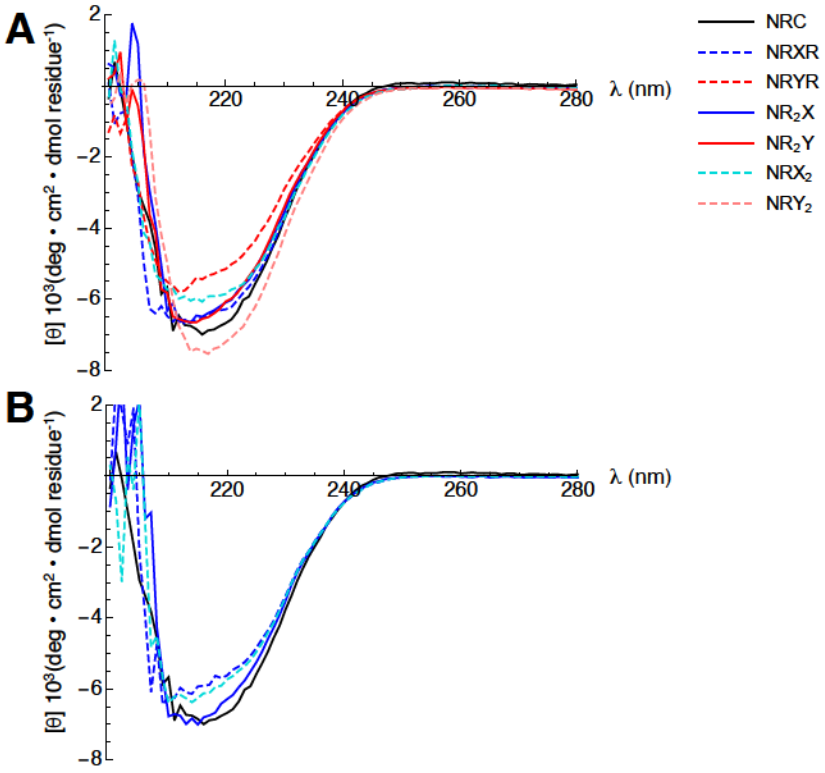


Figure 4.4. Far-UV CD spectra of asparagine ladder cysteine and leucine substitutions. Legend in (A) identifies curves for both (A) and (B). ΔN-cap constructs are not included as concentrations could not be accurately determined for these constructs. (A) CD spectra of $X_{Cys/Leu}/Y_{Cys/Leu}$ constructs for cysteine substitutions. (B) CD spectra of $X_{Cys/Leu}/Y_{Cys/Leu}$ constructs for cysteine substitutions. Solution conditions: 20 mM $NaPO_4$, 500 mM NaCl, 0.1 mM TCEP, pH 7.8, 20 ˚C

4.2.3 Paired repeats parameters for asparagine ladder substitutions.

Unfolding transitions for all of the constructs in Figure 4.3 were collected, and were analyzed with the paired-repeat nearest-neighbor model, using parameters from the original series for ΔG_N, ΔG_R, ΔG_C, and $\Delta G_{R\text{-}1,R}$. Based on intrinsic parameters, both the cysteine and leucine substitutions increase the intrinsic stability of the paired repeat (Table 1). On average, removal of the ladder asparagine reduces the intrinsic folding penalty by about 1.5 kcal mol^{-1} with Y_{Cys} repeats being particularly stabilizing ($\Delta\Delta G$ = -2.2

kcal mol^{-1}, Figure 4.5A). The increased stability of $X_{Cys/Leu}/Y_{Cys/Leu}$ repeats relative to R may result in part from a reduced desolvation penalty of cysteine and leucine, compared to the polar asparagine side chain. Substitution to cysteine and leucine for the ladder asparagine is expected to yield -3.7 kcal mol^{-1} and -4.9 kcal mol $^{-1}$ respectively, though even the most stabilizing substitution (Y_{Cys}) only provides -2.2 kcal mol^{-1} [18]. This discrepancy likely results from the loss of stabilizing hydrogen bonding within the asparagine ladder. In this regard, it is important to keep in mind that X and Y are paired repeats that contain an internal interface (Figure 4.2C). Accounting for the average destabilization of this interface (2.5 kcal mol^{-1}; see below) brings the observed gain in stability into better agreement with that expected from differences in desolvation energies.

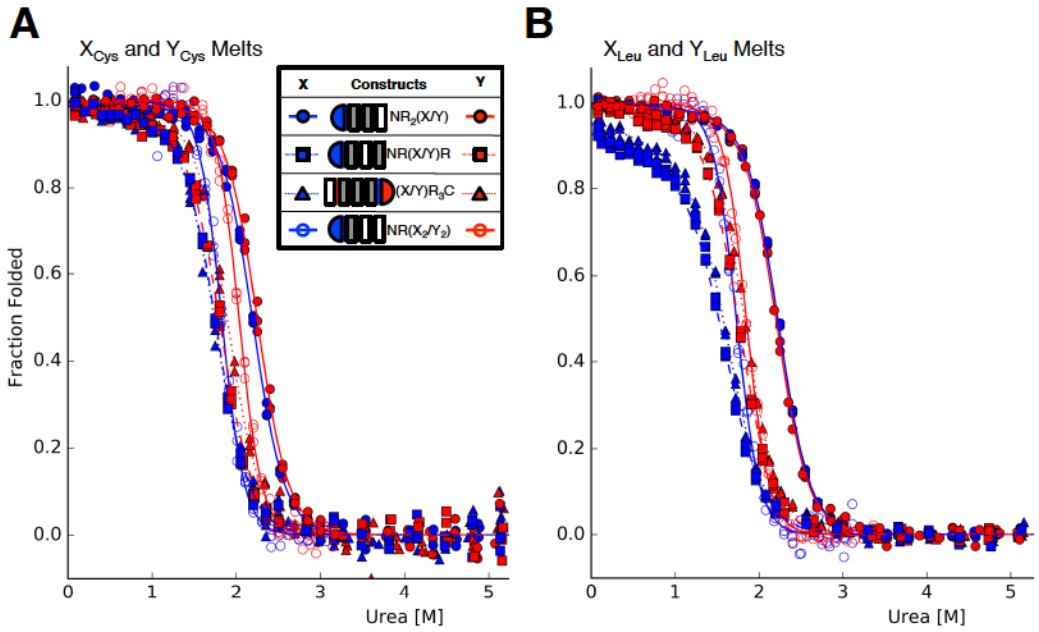


Figure 4.5. Urea-induced unfolding of cysteine and leucine substitutions in cLRR constructs. Plots show fraction folded versus urea concentration with the legend in (A) applying to both plots. Cartoons are as in Figure 4.2. The $(X_{Cys/Leu}/Y_{Cys/Leu})R_3C$ constructs were denatured in 5% glycerol (v/v) to prevent aggregation at low urea concentrations. (A) Urea melts for X_{Cys}/Y_{Cys} substitutions. (B) Fits of urea melts from X_{Leu}/Y_{Leu} constructs. Solution conditions: 20 mM $NaPO_4$, 500 mM NaCl, 0.1 mM TCEP, 0/5% glycerol (v/v), pH 7.8, 20 ˚C.

Conversely, the interfacial free energies are significantly destabilized by disruption of the asparagine ladder (Table 4.1). For each type of substitution, the paired-repeat model gives six interfacial terms: the interfaces N-terminal to the X and Y substitution (R:X and R:Y), the interfaces C-terminal to the X and Y substitutions (X:R and Y:R), and the interfaces between two adjacent substitutions (X:X and Y:Y). All six interfacial parameters have reduced stability relative to the original cLRR interfaces, for both the cysteine and the leucine substitution series. This suggests that neither cysteine nor leucine can replace the favorable interfacial interactions supported by the asparagine ladder.

Table 4.1. Paired repeats parameters for asparagine ladder substitutions.

Asparagine[a]		Cysteine[b]		Leucine[b]	
ΔG_R	10.95	ΔG_X	9.65 [9.04, 10.2]	ΔG_X	9.14 [8.67, 9.54]
		ΔG_Y	8.72 [8.23, 9.17]	ΔG_Y	9.74 [9.52, 10.01]
$\Delta G_{R-1,R}$	-14.2	$\Delta G_{R-1,X}$	-12.06 [-12.26, -11.82]	$\Delta G_{R-1,X}$	-11.53 [-11.80, -11.22]
		$\Delta G_{R-1,Y}$	-11.33 [-11.51, -11.13]	$\Delta G_{R-1,Y}$	-12.06 [-12.16, -11.97]
		$\Delta G_{X-1,R}$	-12.68 [-12.93, -12.47]	$\Delta G_{X-1,R}$	-11.60 [-11.89, -11.28]
		$\Delta G_{Y-1,R}$	-12.65 [-12.81, -12.49]	$\Delta G_{Y-1,R}$	-12.72 [-12.85, -12.61]
		$\Delta G_{X-1,X}$	-11.33 [-12.16, -10.38]	$\Delta G_{X-1,X}$	-10.07 [-10.61, -9.50]
		$\Delta G_{Y-1,Y}$	-10.77 [-11.32, -10.16]	$\Delta G_{Y-1,Y}$	-11.78 [-12.13, -11.46]
m_{urea}	0.74	m_{urea}	0.73 [0.76, 0.70]	m_{urea}	0.72 [0.73, 0.71]

Mean best-fit values from [a]200 or [b]1000 bootstraps iterations. Free energy is in kcal mol^{-1}; m_{urea} is in kcal mol M_{urea}^{-1}. Intrinsic terms are represented as ΔG_i with interfacial terms as $\Delta G_{i-1,i}$ Denaturant dependence is indicated by m_i. 95% confidence intervals for parameters are included in brackets after the mean value. Original cLRR parameters (left column) and glycerol dependence are fixed in all fits with N- and C-terminal caps using 1.5 m_{urea} and 2 m_{urea} respectively to account for the number of repeat pairs in each cap.

Comparing the extent to which the different types of interfaces are destabilized reveals a general hierarchy (Figure 4.6B). Interfaces between two adjacent $X_{Cys/Leu}$ or $Y_{Cys/Leu}$ paired repeats are the most destabilized. Interfaces N-terminal to the site of substitution (R:$X_{Cys/Leu}$ and R:$Y_{Cys/Leu}$) are intermediately destabilized and interfaces C-

terminal to sites of substitution (R:$X_{Cys/Leu}$ and R:$Y_{Cys/Leu}$) are the least destabilized. The exception to this trend is X_{Leu}, where the N- and C-terminal interfaces are equally perturbed. It is surprising that similar stability changes result from A- and B-repeat substitution (i.e., X and Y substitution), suggesting that the effects of ladder disruption can be propagated to nonadjacent interfaces. For example, for the Y_{Cys} series, the substitution in the B repeat is immediately adjacent to the C-terminal interface (Figure 4.2B), yet the N-terminal interfacial parameters is more destabilized than the C-terminal interface (compare cysteine $\Delta G_{R\text{-}1,Y}$ with $\Delta G_{Y\text{-}1,R}$).

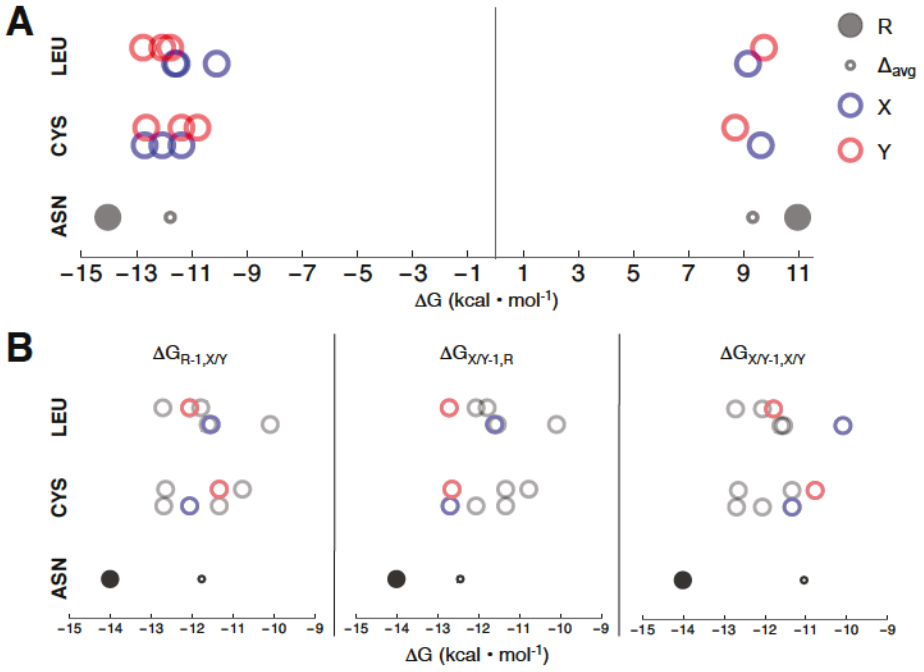


Figure 4.6 Intrinsic and interfacial parameters from nearest-neighbor fits of $X_{Cys/Leu}/Y_{Cys/Leu}$ constructs to the paired-repeats model. $X_{Cys/Leu}$ (blue), $Y_{Cys/Leu}$ (red), and R (gray) parameters are displayed on a free energy number line along the x-axis. Vertical offsets are used to group parameters based on ladder residue identity (asparagine, bottom; cysteine, middle; leucine, top). Δ_{avg} is the average value for all intrinsic or interfacial terms within a plot. (A) Intrinsic and all interfacial parameters for each substitution type. (B) Interfacial terms separated into N-terminal ($\Delta G_{R\text{-}1,X/Y}$), C-terminal, ($\Delta G_{X/Y\text{-}1,R}$), or tandem ($\Delta G_{X/Y\text{-}1,X/Y}$) $(X_{Cys/Leu})_2/(Y_{Cys/Leu})_2$ interface energies. Colored circles are of the parameter type specified by the subplot heading; gray circles represent the other interfacial terms for comparison. Averages in subplots are taken only for colored points.

Although the interfacial parameter hierarchy seems to be independent of which repeat is substituted (A versus B), there are correlations between the type of substitution (cysteine versus leucine) and interfacial stability. The interfaces of X_{Leu} tend to be more destabilized than those of X_{Cys}; however, Y_{Leu} interfaces tend to be less destabilized than those of Y_{Cys} (Figure 4.6B). Interestingly, the stability of the interfaces seems anti-correlated to the stability of the intrinsic terms (Figure 4.6A). Specifically, the asparagine ladder of LRR proteins appears to destabilize folded repeats but stabilizes interfaces between folded repeats, a pattern that promotes cooperativity. Because the effect on interfacial free energies is greater than that on intrinsic folding, the asparagine ladder exerts an overall stabilization on the LRR array.

4.3 Discussion

4.3.1 The role of asparagine ladder in repeat coupling.

The substitution parameters clearly indicate that the asparagine ladder is used as a coupling device in LRR proteins. Both cysteine and leucine substitutions significantly destabilize interfaces on either side of the substituted paired-repeat (Figure 4.6A, Table 4.1). Additionally, the intrinsic terms shed an average of 1.6 kcal mol^{-1} after removal of the ladder asparagine, stabilizing the isolated paired-repeat (Figure 4.6A, Table 4.1). Therefore, the asparagine ladder enhances coupling in two ways: it stabilizes the highly favorable interfaces by 2.5 kcal mol^{-1} on average and simultaneously destabilizes isolated repeats, contributing to the large free energy barrier to isolated repeat folding. This two-

pronged effect is not observed in ankyrin, where the polar ladder greatly stabilizes interfaces but does not destabilize isolated repeats (Figure S4.4) [14].

Anticorrelation between intrinsic and interfacial terms holds for consensus proteins in general, with cLRR substitutions also exhibiting negative correlation between intrinsic and interfacial terms (Figure S4.5). By comparison, ankyrin substitutions show much less deviation in intrinsic stability over a wide range of interfacial stabilities. The insensitivity of the ankyrin intrinsic terms to the stability of the interface appears to set it apart from other consensus systems with substitutions (cLRR and TALEs) or extension of an existing motif (34PR, 42PR). However, the substitution data are still very limited (ankyrins, TALEs, and cLRRs) and the relationship observed in ankyrin may be the norm for most protein systems. Additional substitution studies with the tetratricopeptide-like repeat proteins (34PR and 42PR), TALE, cLRR, or new consensus proteins would better define the relationship between intrinsic and interfacial stability.

One of the most surprising results from this study is the asymmetric destabilization of the interfaces preceding and following a substituted repeat. Analysis of X and Y repeats with both cysteine and leucine substitutions reveals that R:X/Y interfaces are almost always less stable than X/Y:R interfaces (Figure 4.5B). This is particularly notable in the case of Y repeats, where R:Y interfaces are not adjacent to the substitution site but nevertheless are more destabilized than Y:R interfaces ($\Delta G_{Y\text{-}1,R} > \Delta G_{R\text{-}1,Y}$; Table 4.1). It is possible that the molecular architecture of the asparagine ladder is responsible for the asymmetric changes in the interfacial stability between the substituted repeat (*i*) and the preceding ($i - 1^{th}$) and following ($i + 1^{th}$) repeats. Numerous crystal structures show that

the asparagine ladder residue in repeat *i* forms a larger number of hydrogen bonds with the *i* - 1^{th} repeat than to either the i^{th} *or the i +* 1^{th} repeat [12], [19], [20]. Since neither cysteine nor leucine can form the same network of hydrogen bonds as the canonical asparagine, a larger number of bonding interactions to the *i* - 1^{th} repeat are lost. Therefore, substitutions may have a larger effect on the asparagine ladder residue in the *i* - 1^{th} repeat, attenuating the interface between the *i* - 1^{th} and *i* - 2^{th} repeat even though the substitution occurred in the i^{th} repeat. This mechanism would explain the asymmetric destabilization observed in this study, though direct measurement, possibly by monitoring NMR chemical shifts or NOEs, would be required to validate it.

Interfacial asymmetry in substitution effects is not unique to the asparagine ladder and has also been observed in a consensus ankyrin construct (Figure S4.5) [14]. Interestingly, a detailed NMR study of the threonine and histidine residues that form the ankyrin polar ladder revealed that a threonine to valine substitution alters the hydrogen bonding geometry of the histidine [21] and destabilizes the interface between the substituted repeat and the following repeat [14] despite threonine making no direct interactions with the following repeat. In conjunction, these data suggest that effects from substitutions can propagate through a hydrogen bond network, providing an example of a mechanism that may be at work in the asparagine ladder of LRR proteins. Additionally, the consensus ankyrin and cLRR studies quantify the relationship between interfacial stability and the distributions of interactions providing an excellent example of an intuitive result: the greater the number of contacts between two residues, the larger the destabilization will be upon their substitution.

The asparagine ladder's asymmetric stability is particularly important when considering how it contributes to global cooperativity. As the asparagine ladder couples repeats, substitutions to the asparagine ladder are expected to reduce cooperativity and populate partially folded states. The nearest-neighbor model parameters can be used to reconstruct the population of partially folded states as a function of denaturant concentration (Figure 4.7), showing that the population of partially folded states is highly dependent on the substitution identity and location. In some cases, the presence of partially folded states is significant without any denaturant (Figure 4.7, top right), explaining the highly sloped native baselines in the nearest-neighbor fits of some X_{Leu} constructs (Figure 4.5B). The impressive sensitivity of the population of partially folded states to substitution type and location may be biologically relevant in natural LRRs with broken asparagine ladders.

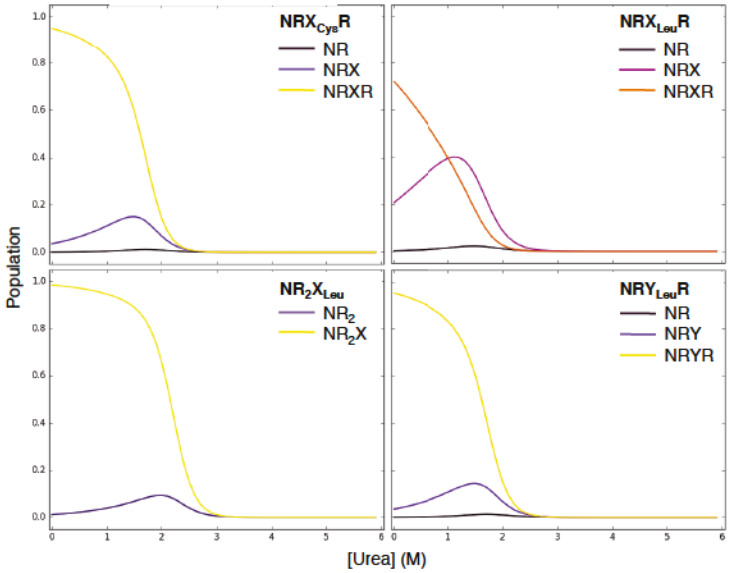


Figure 4.7. Population of partially folded states in substituted cLRR constructs. Plots of fully- or partially-folded states (populated ≥ 1% during chemical denaturation) for four cLRR constructs. Legends in the top right corner of each plot indicate the construct and fully-/partially-folded states present during chemical denaturation. Line colors reflect population of each state at 0 M urea.

An example of a biological interaction that may take advantage of the asparagine ladder to manipulate partially folded states is the interaction between the LRR domain of pp32 and the nuclear transporter Crm1 [22]. A previous study hypothesized that the biased stability of pp32 ($\Delta G_{N\text{-}term} > \Delta G_{C\text{-}term}$) may be required for exposure of a cryptic nuclear export signal for interacting with Crm1 [7]. Interestingly, the first two LRRs in pp32 lack asparagines in the ladder positions, which could lead to substantial population of states where the first two LRRs of pp32 are unfolded (Figure 4.5 and Figure 4.7, top right). It may be that pp32 lacks an asparagine ladder in its N-terminal LRRs to promote local unfolding, exposing nuclear export sequence(s) that would otherwise be sequestered in LRR secondary structural elements.

4.3.2 Coupling between asparagine ladder positions.

The paired-repeats parameters can be used to derive the long-range nonadditivity between the *i* and *i* ± *2*th interfaces in the $(X_{Cys/Leu})_2$ /$(Y_{Cys/Leu})_2$ constructs (Figure S4.6) [23]. For all substitutions but X_{Cys} coupling exceeds the error propagated from the 95% confidence interval of fitted parameters, indicating significant coupling between non-adjacent asparagine ladder positions. Though error is high relative to the size of the couplings, a coupling of 1 kcal mol^{-1} exists between ladder positions in repeat *i* and *i* ± *2* (Table 4.1). Although coupling between adjacent repeats was not measured in this study, it has been shown that couplings can be as high as -5 kcal mol^{-1} in the LRR protein pp32 (see Chapter 2).

It is informative to compare the coupling between asparagine ladder positions in the cLRR protein (and the pp32 proxy) to those in other consensus proteins. For TALEs, no significant couplings exist between adjacent repeat variable diresidue positions (Table S4.1). For ankyrin, strong coupling is observed between two positions, one of which is part of a threonine-histidine polar ladder (T4V; Figure S4.1) [21]. Threonines in ankyrin's polar ladder are separated by a histidine residue, so this coupling is between positions that do not directly interact. Therefore, both ankyrin and cLRR systems have long-range coupling within their polar ladder, demonstrating the strong energetic interconnectivity within polar ladders generally.

4.3.3 Comparison of probable and improbable substitutions

As mentioned in the results section, the choice of cysteine and leucine substitutions was based on a relationship between frequency and stability; cysteine, a more frequent ladder substitution, was expected to be more stable than leucine, which is observed less frequently at the ladder position. Free energies derived from two-state fits to cLRR $X_{Cys/Leu}$ and $Y_{Cys/Leu}$ constructs show that cysteine is generally less destabilizing than leucine for a given construct (Table S4.2), consistent with the idea that frequencies are related to stability.

The global stabilities of the substituted cLRR constructs must be partitioned into the paired-repeats parameters. Since the asparagine ladder contributes to coupling between repeats, the stability difference between cysteine and leucine at the ladder position might be expected to partition into the interfacial terms. However, the interfacial

terms do not exhibit a strong stability bias towards either substitution (Figure S4.7A). Only when the intrinsic and interfacial terms are considered together does cysteine appear to be less destabilizing (Figure S4.7B). This finding provides further evidence that neither cysteine nor leucine can fully recapture the interfacial interactions supported by asparagine at the ladder position. Cysteine may be less destabilizing than leucine because it is smaller, making it easier to pack into the cLRR hydrophobic core (reducing the intrinsic cost of folding) while maintaining as much of the favorable interface as possible. A firmer conclusion would require a high-resolution structure of both mutants, which could identify potential steric clashes or rearrangements that might favor one substitution over the other.

The greater stability of cysteine relative to leucine substitutions appears to support the argument that the asparagine ladder in LRR proteins is independent of sequence context. Despite being no more common than leucine, cysteine exhibits a small but significant stability increase. The differences between the two datasets may result from sampling, as the full LRR data set had two orders of magnitude more sequences than the bacterial LRR one. If true, the stability of substitutions to the invariant LRR region in the cLRR sequence might be better described by the conservation patterns in the full LRR dataset. This hypothesis could be tested by substituting threonine at the asparagine ladder given threonine is observed twice as frequently as cysteine in the bacterial LRR MSA but one fifth as much as cysteine in the full LRR MSA.

4.4 Materials & Methods

4.4.1 Protein cloning and expression.

Proteins were expressed and purified as in Chapter 3.

4.4.2 Circular dichroism spectra and equilibrium unfolding.

All CD experiments were performed on an Aviv Model 400 CD spectrometer using a computer-controlled Microlab syringe titrator (Hamilton) with samples in CD buffer (20 mM Na PO4, 500 mM NaCl, 0.1 mM TCEP, 0/5% glycerol (v/v), pH 7.8) at 20 ˚C. CD spectra were collected with 3 s signal averaging every nm from 280 to 200 nm; protein concentrations were 30 to 50 μM in a 0.1 mm path-length quartz cuvette. Equilibrium unfolding experiments were monitored at 220 nm using a 5 to 6 min equilibration time and 30 s signal averaging; protein concentrations were 3-5 μM in a 1 cm path length quartz cuvette. Urea (VWR Life Sciences) used for experiments was first deionized using a mixed-bed resin (BioRad) and concentrations were determined from the refractive index [24]. Glycerol dependence for repeats was determined from melts of NR_2 and NRC constructs denatured in 0, 2.5, and 5% glycerol (v/v) (see Chapter 3). Unfolding experiments with ΔN-cap constructs were all done with 5% glycerol (v/v), titrating from high urea concentrations to lower concentrations with no protein in the titrant to avoid aggregation of titrant protein at low urea concentrations over the course of the experiment. For the $X_{Leu}R_3C$ construct, melt conditions were contaminated with 10 mM arginine used during dialysis for that prep only. This did have a measurable effect on the CD signal that did not interfere with the collection of the titration data but the low

concentration of arginine is not expected to alter the stability of the construct and thus it was fit the same as the other ΔN-cap constructs.

4.4.3 Nearest-neighbor analysis.

Determination of single repeat interfacial and intrinsic parameters was achieved using a 1D nearest-neighbor model [25]. The parameters for original cLRR repeats ($\Delta G_{N/R/C}$, $\Delta G_{R\text{-}1,R}$) and glycerol dependence (m_{glyc}) were fixed using the values from chapter 2. In addition to the fixed original cLRR parameters, substituted parameters were introduced:

$$\kappa_X = e^{-(\Delta G_X - m_i * x - m_g * g)\beta} \quad (4.1)$$

$$\kappa_Y = e^{-(\Delta G_Y - m_i * x - m_g * g)\beta} \quad (4.2)$$

$$\tau_{R-1,X/Y} = e^{-\Delta G_{R-1,X/Y}\beta} \quad (4.3)$$

$$\tau_{X/Y-1,R} = e^{-\Delta G_{X/Y-1,R}\beta} \quad (4.4)$$

$$\tau_{X/Y-1,X/Y} = e^{-\Delta G_{X/Y-1,X/Y}\beta} \quad (4.5)$$

where β is $(kT)^{-1}$, x is urea concentration in molar, and g is glycerol concentration in molar. As no construct has an $X_{Cys/Leu}/Y_{Cys/Leu}$ adjacent to an N- or C-cap, no unique interfacial terms or assumptions were needed for the nearest-neighbor analysis. Using these equilibrium constants, a partition function can be constructed and used to determine the fraction of repeats folded in the array (see Introduction) [25]. Parameters were determined using nonlinear least squares to globally fit normalized urea denaturation data [26] using

a script designed by Dr. Jake Marold, modified by Dr. Katie Geiger-Schuller, and with additional minor modifications.

4.5 Supplementary Figures & Tables

Table S4.1. Thermodynamic couplings between positions in *i* and *i ± 1* or *i ± 2* repeats in consensus proteins.

cLRR		Ankyrin[a]		TALE[b]	
$\Delta G_{XCys,i+2}$	-0.6 ± 1.0	$\Delta G_{T4V,i+1}$	-1.8 ± 0.4	ΔG_{i+1}	-0.5 ± 0.8
$\Delta G_{XLeu,i+2}$	-0.9 ± 0.7	$\Delta G_{L6I,i+1}$	-0.5 ± 0.5		
$\Delta G_{YCys,i+2}$	-0.8 ± 0.6	$\Delta G_{L21F,i+1}$	0.1 ± 0.3		
$\Delta G_{YLeu,i+2}$	-1 ± 0.5	$\Delta G_{V28P,i+1}$	-1.9 ± 0.5		

Thermodynamic couplings derived from double-mutant cycles in three consensus proteins as demonstrated in Figure S4.4. Free energy is in kcal mol^{-1} and terms are represented as $\Delta G_{j,i+n}$ where *j* is the identity of the substitution (only one possible in TALEs) and *n* is the number of interfaces from substitution *i*. Errors are propagated from 95% confidence intervals from bootstrap iterations. [a] Data from [14]. [b] Data from [13].

Table S4.2. Comparison of two-state fits of $X_{Cys/Leu}$/$Y_{Cys/Leu}$ constructs.

	Asparagine	Cysteine	Leucine
NR_3	-6.99 ± 0.4		
NR_2X		-7.3 ± 0.4	-7.4 ± 0.1
NR_2Y		-7.4 ± 0.3	-7.5 ± 0.1
$NRXR^a$		-6.0 ± 0.4	-4.8 ± 0.0
$NRYR^a$		-6.5 ± 0.0	-5.8 ± 0.2
R_4C	-5.35 ± 0.1		
XR_3C^c		-6.5 ± 0.5	-6.1 ± 1.4
$YR_3C^{a,c}$		-8.2 ± 0.1	-6.8 ± 0.3
NRX_2		-4.9 ± 0.3	-4.9 ± 0.0
NRY_2^b		-4.7 ± 0.6	-5.7 ± 0.1

ΔG°_{H2O} parameters from two-state fits of $X_{Cys/Leu}$ and $Y_{Cys/Leu}$ constructs. Values are the mean of three titrations with error equal to the standard deviation between titrations. Free energy is in kcal mol^{-1}. [a] Constructs where cysteine is more stable than leucine. [b] Constructs where leucine is more stable than cysteine. [c] Constructs measured with 5% glycerol (v/v). Conditions: 20 mM $NaPO_4$, 500 mM NaCl, 0.1 mM TCEP, 0/5% glycerol (v/v), pH 7.8, 20 °C.

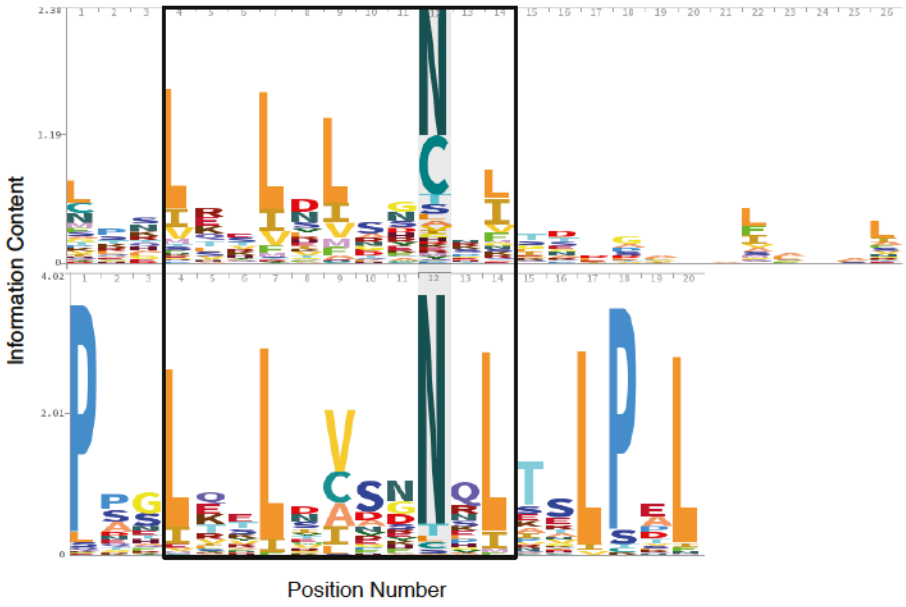


Figure S4.1. Conservation patterns in general LRR versus bacterial LRR subfamily conservation. Comparison of LRR conservation patterns (top) used for choosing cLRR substitutions and those in bacterial LRR (bottom) used to design the cLRR sequence. The invariant LRR region is bounded by the black rectangle. The conserved asparagine position is highlighted in gray. The LRR web logo (top) is longer, reflecting inclusion of LRR classes composed of more residues (typical, RI, etc.). The aggregated LRR web logo differs from the logo in Figure 4.1 as the first two residues in that sequence logo were moved to the end to better align with the cLRR sequence in the figure.

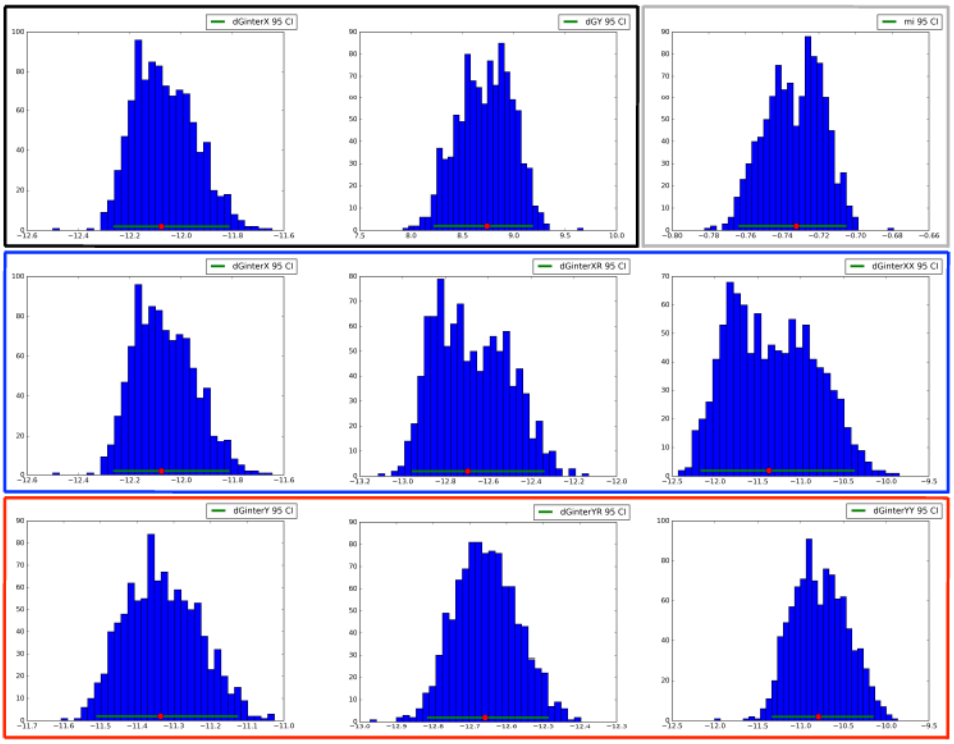


Figure S4.2 Uncertainty in parameters from paired-repeat model of cysteine substitutions. Parameter value distributions from 1000 bootstrap iterations of the paired-repeats nearest-neighbor model for the cysteine substitution. Intrinsic (black box), urea and glycerol dependence (gray box), X_{Cys} interfacial terms (blue box), and Y_{Cys} interfacial terms (red box) parameter values are shown as histograms with the parameter name in the top right of each subplot. Mean value is shown by a red point; 95% confidence intervals are shown in green. Units for subplots are kcal mol^{-1} (intrinsic, black; interfacial, blue and red) and kcal mol^{-1} M^{-1}_{urea} (m_i, gray).

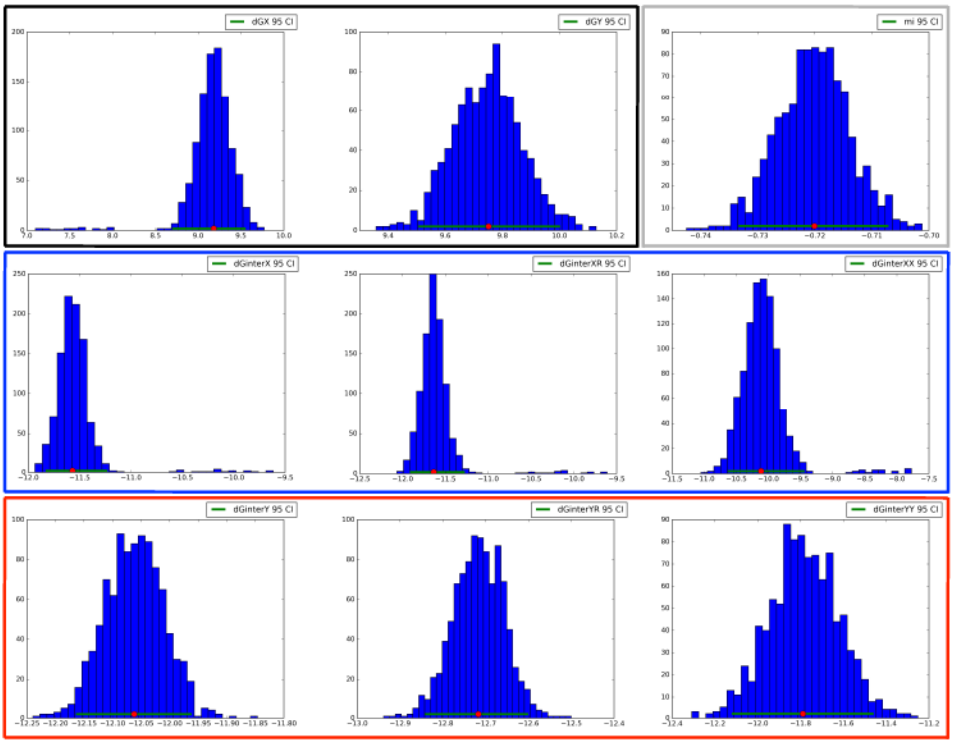


Figure S4.3 Parameter value distributions from 1000 bootstrap iterations of the paired-repeats nearest-neighbor model for the leucine substitution. Intrinsic (black box), urea and glycerol dependence (gray box), X_{Leu} interfacial terms (blue box), and Y_{Leu} interfacial terms (red box) parameter values are shown as histograms with the parameter name in the top right of each subplot. Mean value is shown by a red point; 95% confidence intervals are shown in green. Units for subplots are kcal mol^{-1} (intrinsic, black; interfacial, blue and red) and kcal mol^{-1} M^{-1}_{urea} (m_i, gray).

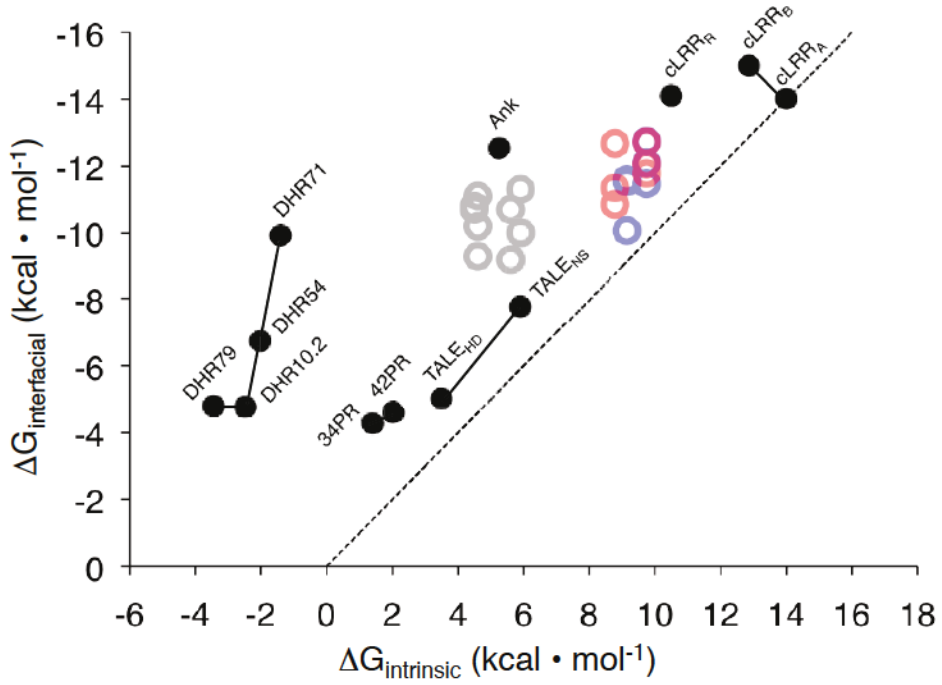


Figure S4.4. Correlation between interfacial and intrinsic ΔG values for consensus proteins. All un-substituted consensus constructs are labeled and proteins from similar families are joined by line. Values for point substitutions are shown in circles and are colored by family (gray, ankyrin; red/blue, cLRR). The dashed line represents a slope of one. A linear fit of the data has slope -0.5 and $R^2 = 0.666$.

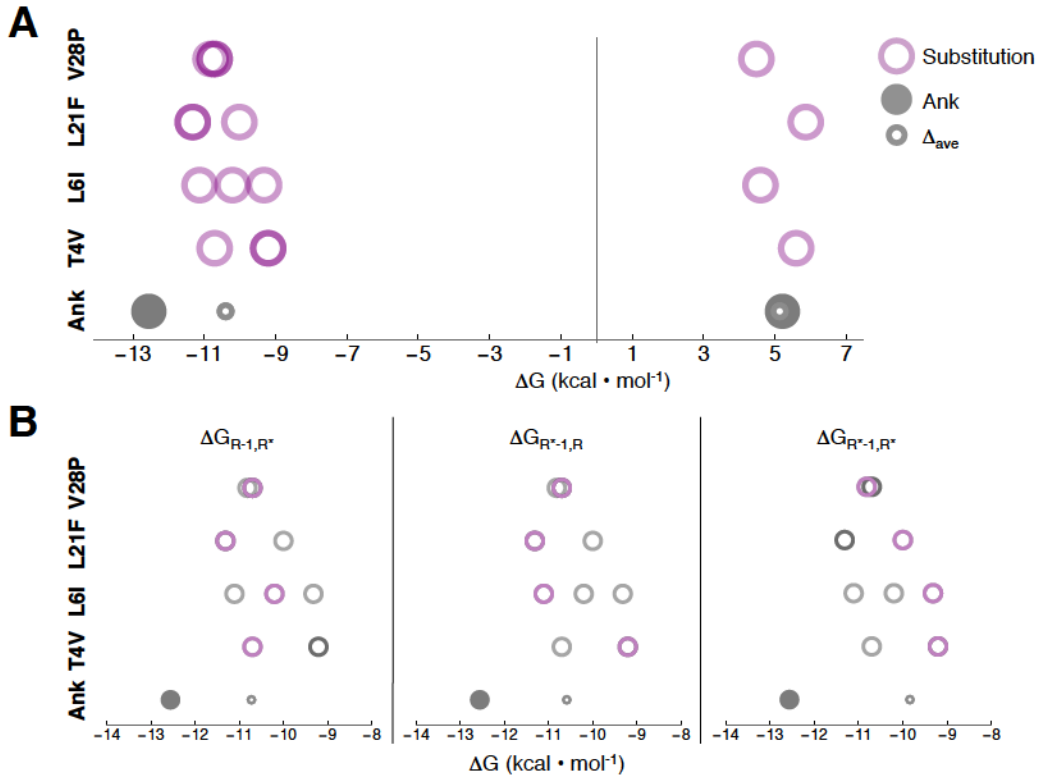
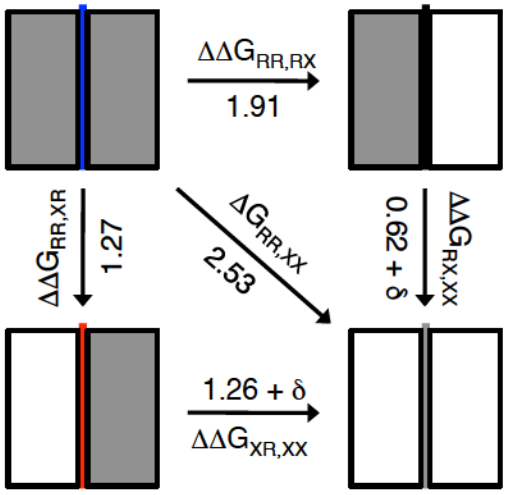


Figure S4.5 Intrinsic and interfacial parameters from nearest-neighbor fits of ankyrin variants. Substituted (purple circles) and unsubstituted (gray) parameters are displayed on a free energy number line along the x-axis. Vertical offsets are used to group parameters based on substitution. Δ_{avg} is the average value for all intrinsic or interfacial terms within a plot. (A) Intrinsic and all three interfacial parameters for consensus ankyrin and for each substitution. The average intrinsic term overlaps the unsubstituted ankyrin consensus term. (B) Interfacial free energies separated into N-terminal ($\Delta G_{R\text{-}1,R^*}$), C-terminal, ($\Delta G_{R^*\text{-}1,R}$), and tandem ($\Delta G_{R^*\text{-}1,R^*}$) interface values. Colored circles are of the parameter type specified by the subplot heading; gray circles represent the other interfacial terms for comparison. Averages in subplots are taken only for colored points.



$$\Delta\Delta G_{RR,XX} = \Delta\Delta G_{RR,RX} + \Delta\Delta G_{RR,XR} + \delta$$
$$\delta = \Delta\Delta G_{RR,XX} - \Delta\Delta G_{RR,RX} - \Delta\Delta G_{RR,XR}$$

Figure S4.6. Double-mutant cycle for X_{Cys}. Original cLRR paired repeats (gray bars) are converted to X_{Cys} paired repeats (white bars) changing R:R interfaces (blue line) to R:X_{Cys} (black line), X_{Cys}:R (red line), and X_{Cys}:X_{Cys} (gray line) interfaces. Transitions are labeled with the ΔΔG from the nearest-neighbor models of X_{Cys}. The equations (bottom) indicate how coupling between the ladder positions of repeats *i* and *i + 2* (δ) is calculated.

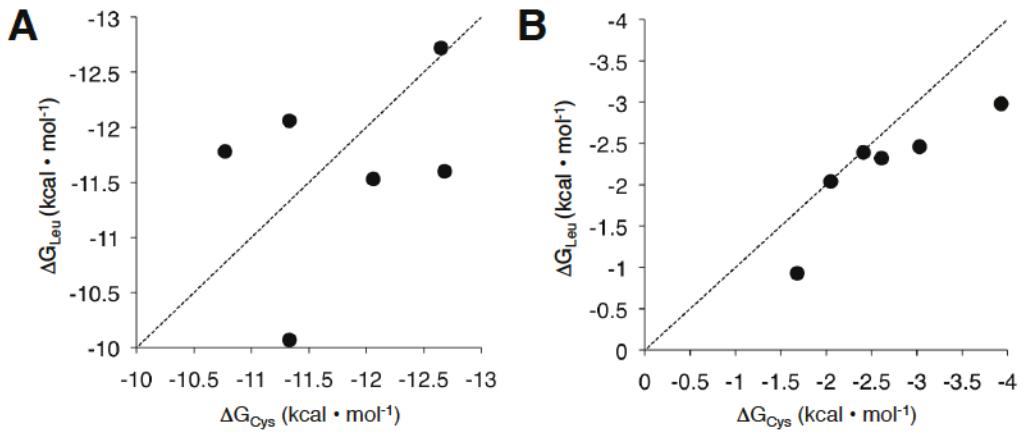


Figure S4.7. Correlation between interfacial parameters in cLRR substitutions. Correlation of interfacial (A) and interfacial + intrinsic (B) terms from cysteine and leucine substitutions with substitution identity. Dashed lines represent unit slope. (A) Correlation between interfacial parameters from cysteine and leucine substitutions. (B) Correlations between intrinsic + interfacial terms for cysteine and leucine substitutions. Points represent $\Delta G_{X/Y} + \Delta G_i$, where ΔG_i is one of the three interfacial terms for each substitution.

4.6 References

[1] C. O. Mackenzie and G. Grigoryan, “Protein structural motifs in prediction and design,” *Current Opinion in Structural Biology*. 2017.

[2] P. S. Huang, S. E. Boyken, and D. Baker, “The coming of age of de novo protein design,” *Nature*, vol. 537, no. 7620, pp. 320–327, 2016.

[3] M. J. Harms and J. W. Thornton, “Historical contingency and its biophysical basis in glucocorticoid receptor evolution,” *Nature*, 2014.

[4] E. Kloss, N. Courtemanche, and D. Barrick, “Repeat-protein folding: New insights into origins of cooperativity, stability, and topology,” *Arch. Biochem. Biophys.*, vol. 469, no. 1, pp. 83–99, 2008.

[5] N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, and R. Sun, “Adaptation in protein fitness landscapes is facilitated by indirect paths,” *Elife*, 2016.

[6] V. H. Salinas and R. Ranganathan, “Coevolution-based inference of amino acid interactions underlying protein function,” *Elife*, 2018.

[7] T. P. Dao, A. Majumdar, and D. Barrick, “Highly polarized C-terminal transition state of the leucine-rich repeat domain of PP32 is governed by local stability,” *Proc. Natl. Acad. Sci.*, 2015.

[8] E. Kloss and D. Barrick, “Thermodynamics, Kinetics, and Salt dependence of Folding of YopM, a Large Leucine-rich Repeat Protein,” *J. Mol. Biol.*, vol. 383, no. 5, pp. 1195–1209, 2008.

[9] L. K. Mosavi, D. L. Minor, and Z. -y. Peng, “Consensus-derived structural determinants of the ankyrin repeat motif,” *Proc. Natl. Acad. Sci.*, 2002.

[10] T. Kajander, A. L. Cortajarena, E. R. G. Main, S. G. J. Mochrie, and L. Regan, “A new folding paradigm for repeat proteins,” *J. Am. Chem. Soc.*, 2005.

[11] T. Aksel and D. Barrick, “Direct observation of parallel folding pathways revealed using a symmetric repeat protein system,” *Biophys. J.*, vol. 107, no. 1, pp. 220–232, 2014.

[12] J. Bella, K. L. Hindle, P. A. McEwan, and S. C. Lovell, “The leucine-rich repeat structure,” *Cell. Mol. Life Sci.*, vol. 65, no. 15, pp. 2307–2333, 2008.

[13] K. Geiger-Schuller and D. Barrick, “Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States,” *Biophys. J.*, vol. 111, no. 11, pp. 2395–2403, 2016.

[14] K. A. Sforza, *Alpha-helical repeat protein folding and turnover: A thermodynamic analysis of natural and unnatural repeat architectures*. Johns Hopkins University, 2017.

[15] W. M. Fitch and E. Markowitz, “An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution,” *Biochem. Genet.*, 1970.

[16] B. Kobe and A. V. Kajava, “The leucine-rich repeat as a protein recognition motif,” *Current Opinion in Structural Biology*. 2001.

[17] E. Kloss and D. Barrick, “C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein,” *Protein Sci.*, vol. 18, no. 9, pp. 1948–1960, 2009.

[18] D. C. Marx and K. G. Fleming, “Influence of Protein Scaffold on Side-Chain Transfer Free Energies,” *Biophys. J.*, 2017.
[19] A. G. Evdokimov, D. E. Anderson, K. M. Routzahn, and D. S. Waugh, “Unusual molecular architecture of the Yersinia pestis cytotoxin YopM: A leucine-rich repeat protein with the shortest repeating unit,” *J. Mol. Biol.*, 2001.
[20] B. Kobe and J. Deisenhofer, “Proteins with leucine-rich repeats,” *Curr. Opin. Struct. Biol.*, 1995.
[21] M. R. Preimesberger *et al.*, “Direct NMR detection of bifurcated hydrogen bonding in the α-helix N-caps of ankyrin repeat proteins,” *J. Am. Chem. Soc.*, 2015.
[22] C. M. Brennan, I. E. Gallouzi, and J. A. Steitz, “Protein ligands to HuR modulate its interaction with target mRNAs in vivo,” *J. Cell Biol.*, 2000.
[23] A. Horovitz, “Double-mutant cycles: A powerful tool for analyzing protein structure and function,” *Fold. Des.*, vol. 1, no. 6, pp. 121–126, 1996.
[24] C. N. Pace, “Determination and Analysis of Urea and Guanidine Hydrochloride Denaturation Curves,” *Methods Enzymol.*, 1986.
[25] T. Aksel and D. Barrick, *Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models*, 1st ed., vol. 455, no. A. Elsevier Inc., 2009.
[26] K. Geiger-Schuller, K. Sforza, M. Yuhas, F. Parmeggiani, D. Baker, and D. Barrick, “Extreme stability in de novo-designed repeat arrays is determined by unusually stable short-range interactions,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 29, p. 201800283, 2018.

CHAPTER 5 – Future directions

5.1 Discussion

This thesis has explored many features of the asparagine ladder and has provided some direction for future studies into the asparagine ladder in leucine-rich repeat (LRR) proteins. The most favorable avenues to explore are listed below with some thought given to what question each might address and potential challenges associated with the necessary experiments. Though this is not an exhaustive list, it provides a good starting point for future investigations.

5.1.1 Hydrogen exchange in extended asparagine ladders.

The initial studies of the asparagine ladder in pp32 determined that it is highly protected from solvent (see Chapter 2). However, the short length of the ladder and the biased stability of pp32 prevented a more detailed investigation into whether protection increases with increasing distance from the end of the ladder. Fortunately, both YopM and cLRRs have extensive ladders that might be able to resolve the protection gradient from the ladder's end to its center. An important consideration is that both these systems are considerably larger than pp32 (385 residues for YopM and 216 residues for NR_2 versus 157 residues), which is likely to exacerbate the issues with weak signal from ladder asparagine side chains that arose in the pp32 protein. However, YopM truncations of up to 4-6 repeats are relatively stable [1] and would provide a significant enhancement in tumbling time relative to the full-length protein (239 residues after truncating C-terminal 6

repeats). The consensus construct NR_2 is both more stable and smaller than the smallest stable YopM truncation, possibly making it a more attractive target. However, if the asparagine ladder between the consensus repeats and the N-terminal cap is broken, the size of the resolvable ladder in the NR_2 construct is significantly smaller than that of YopM truncations.

5.1.2 Long-range coupling studies of asparagine ladder.

Although *i* to *i±2* couplings were resolved in the cLRR system in Chapter 4, it would be worthwhile to see if any coupling exists beyond two repeats from the substitution site. Though this kind of study could be done in the cLRR system, creating the necessary constructs would be more challenging than mutagenesis in a natural protein like YopM. Data from the cLRR study show that couplings between *i* and *i±2* ladder positions are small, so initial experiments to determine if this coupling can be measured in the YopM system would be a necessary first step. After determining if long-range coupling can be measured in the YopM system, further experiments could be conducted to measure even further couplings or test how varied coupling is for different sites within YopM. This may be particularly interesting to measure for internal-capping repeats [1].

5.1.3 High-resoultion structures of cLRR constructs.

Though studies of the cLRR protein (see Chapter 3 and 4) indicate that it possesses the characteristic LRR structure, a high-resolution crystal structure would clarify some of the unanswered questions from these previous studies. Substantial effort has already been spent on crystalizing a number of cLRR constructs, including substituted variants. Thus far, the NR_3C construct has shown the most promise as it has crystallized in a number of different conditions. Unfortunately, the initial crystals diffracted poorly and efforts to optimize crystallization conditions were unsuccessful. Further efforts to remove unstructured regions are currently underway and will hopefully improve the quality of crystals. In addition, fusion constructs with maltose-binding protein might be used to further improve the chances of generating high-resolution crystals of a cLRR construct [2].

Although the NR_3C construct is the most promising hit, it would be extremely valuable to obtain structures of constructs with asparagine ladder substitutions. This might provide some insights into the structural effects of broken ladders. Furthermore, it may be the case that structures do not exist for some ladder architectures, such as the threonine and cysteine ladders that are hinted at in the sequence coupling data in Chapter 3. The cLRR system may be a useful scaffold upon which to do structural studies of the effects of these substitutions in single repeats or in arrays.

5.1.4 Single-repeat model for substitutions in cLRR constructs.

The paired-repeat model used to evaluate substitutions in chapter 4 provided a strong foundation to study the effects of substitutions in the cLRR system. It would be useful to extend this work by performing melts with substitutions to the necessary constructs for the single-repeats analysis in chapter 3 (NR_3A and BR_3C). This would help disentangle any sequence-dependent effects that might arise from the charge alternating substitutions in the A and B repeats. Unfortunately, it is likely that substituted NR_3A constructs would also remain unfolded, making it impossible to resolve the full set of single-repeat parameters. It may be prudent to see if redesigning the charge alternating residues, particularly the aspartate at position ten (see Chapter 3), using the mutual information data from the bacterial LRR subfamily. This might improve the stability of the A repeat and B:A interfaces, allowing for complete resolution of the single-repeat parameters for the unsubstituted and substituted cLRR constructs.

5.2 References

[1] E. Kloss and D. Barrick, "C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein," *Protein Sci.*, vol. 18, no. 9, pp. 1948–1960, 2009.

[2] D. S. Waugh, "Crystal structures of MBP fusion proteins," *Protein Science*. 2016.

Curriculum Vitae

Sean Klein was born in Ames, Iowa on June 29, 1989 to parents Ron and Michelle Klein. Shortly thereafter, he moved with his family to Peoria, where his brother Nathan Klein was born, and then Morton, Illinois. After attending Morton Public High School, Sean began his undergraduate education at the University of Iowa. Along with the coursework for his biochemistry degree, Sean explored a number of history, art, philosophy, and law classes to round out his college education. His undergraduate research experience began his sophomore year in the lab of Charles Brenner, working with Jennifer Boylston. In his senior year and at the suggestion of his friend and classmate Sterling Martin, Sean moved to a new research lab working with Madeline Shea on a project involving calmodulin interactions with calcineurin.

After graduating in 2012 with his B.S. in Biochemistry, Madeline generously allowed Sean to continue his research as a paid post-baccalaureate researcher. In Madeline's lab, Sean met lifelong friends and, inspired by them and Madeline, decided to pursue a PhD from the Program in Molecular Biophysics at Johns Hopkins University. At Hopkins, Sean rotated in the labs of Jin Zhang, James Berger, and Doug Barrick; he ultimately selected the Barrick lab for his thesis research studying the thermodynamics and cooperativity of leucine-rich repeat proteins. After six incredible years with Doug and his lab, Sean completed his dissertation research in the summer of 2019.